DOCUMENT RESUME

ED 249 271 TM 840 627

AUTHOR	Hills, John R.
TITLE	Quantitative Methods Used in the Study of Item Bias.
SPONS AGENCY	Florida State Dept. of Education, Tallahassee.
PUB DATE	[84]
NOTE	165p.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Evaluation Methods; Item Analysis; *Racial Differences; *Test Bias; Test Construction; *Testing Problems; Test Items; Test Results

ABSTRACT

The literature on item bias, i.e., the question of whether some items in tests favor one cultural group over another cultural group due to irrelevant factors, is reviewed and evaluated. All known references through 1981 are described including a large number of unpublished reports. Each method is described and the criticisms that have appeared in the literature are reported. Studies evaluating methods, comparing methods, and using the methods in practical applications are described and criticized. In the two decades of work on this topic, little has been accomplished that will result in reduction in the mean differences between cultural groups. The probability seems high that well designed and developed modern tests contain few items biased in the sense defined above. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED249271

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
J. R. Hills

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Quantitative Methods Used in the Study of Item Bias

by

John R. Hills

Florida State University

Correspondence concerning this article should be directed to the author, Room 307 Stone Building, Florida State University, Tallahassee, Florida, 32306.

This literature review was partially supported by contracts from the Florida Department of Education. In the early stages, Dr. Nicholas Maruhnich was a collaborator. Dr. William Angoff reviewed an earlier version of the manuscript and made many helpful suggestions. This assistance is greatly appreciated. However, the views expressed are solely those of the author.

TM 810 627

Quantitative Methods Used in the Study of Item Bias

Abstract

The literature on item bias, i.e., the question of whether some items in tests favor one cultural group over another cultural group due to irrelevant factors, is reviewed and evaluated. All known references through 1981 are described including a large number of unpublished reports. Each method is described and the criticisms that have appeared in the literature are reported. Studies evaluating methods, comparing methods, and using the methods in practical applications are described and criticized. In the two decades of work on this topic, little has been accomplished that will result in reduction in the mean differences between cultural groups. The probability seems high that well designed and developed modern tests contain few items biased in the sense defined above.

Introduction

Bias is a problem in educational testing. It has received extensive attention during the last decade. Bias can be defined generally as deviation from truth, i. e., a score on a test or on an item, or a procedure for using a test score, that systematically misleads or deviates from that which is correct (Angoff, Note 1, p. 2). The issue of item bias stems from the fact that for many years it has been observed repeatedly that different cultural groups obtain different mean scores on mental tests. The fact that often in the United States black and Hispanic examinees on the average score markedly lower than white examinees has led to the charge that these differences in mean scores are due to the fact that items on the tests are biased in favor of white examinees.

The fact of a difference in mean scores between groups does not, by itself, prove that items on a test are biased. There may be real differences between the groups that are reflected in the test scores. To say that an item is biased because a smaller proportion of group X than group Y answers it correctly would be equivalent to a doctor's deciding that his scale for measuring weight was biased because undernourished children weigh less than adequately nourished children. Obviously, the doctor would say, "Feed the undernourished children" instead of saying, "Let's correct their weights for the bias in the scale." (Similar arguments could be developed around the fact that women's feet are typically shorter than men's feet, or the fact that the blood pressure of blacks is typically higher than that of whites.) If the measurements are from mental tests and the differences between groups are real differences, the educator might well argue that the children in the low-scoring group should be helped to develop to a higher level of performance. Thus, when Breland and his co-workers (Breland, et al., Note 2, pp 49-54) found as a result of searching for biased items that minority children had much more difficulty than middle-class children in deciding whether the square root of 9 was larger or smaller than 9, the Breland group was reluctant to say that children should no longer be asked this question. Instead, it was inclined to suggest that more effective efforts to teach minority children simple quantitative concepts were needed.

If the items are not biased, and the differences between groups are real, then how could those differences have arisen? One possibility is that there are genetic differences between the groups, as in the case of men's and women's length of foot. Another possibility is that there are cultural differences between the groups. Possible cultural differences are such things as children from different cultural backgrounds may have different levels of motivation and interest for learning the abstract and sometimes seemingly useless things taught in school, such as the square root concept. Maybe students from different cultural groups are taught different things. Maybe they learn different things outside of formal school instruction. Note, however, that if items detect real differences due to such causes as these, the items are not biased. They are doing exactly what good items should do--as long as the element which causes different levels of group performance on the item is relevant to the objectives of the curriculum.

For example, consider again the item noted by Breland, "Which is larger, 9 or the square root of 9?" Let's assume that knowledge of the square root concept was one of the objectives of the curriculum for these children. One cannot easily determine what about the question, "Which is larger, 9 or the square root of 9?" could be irrelevant to the instructional objective and causing the difference in performance level.

One might conjecture that the school taught white students the concept but did not teach it to black students. That seems unlikely, but even if it were true, the data do not indicate that the item is biased; they indicate that the instruction was biased. The item should not be eliminated as a faulty item, but should be retained for what it reveals about instruction.

It seem unlikely that instruction would be deliberately biased in such a way in a school as such schools are now constituted in the United States. Perhaps, then, the black children just were not interested in learning square root when they were instructed in it. Again, the item is not biased. Nothing is irrelevant about it that causes the difference in performance. In fact, the item would be revealing, then, an important fact about student motivation. Thus, one cannot validly assert item bias if the missing knowledge is the result of different children being taught, or learning, different things (Coffman, 1978). Lack of knowledge is not a valid basis for asserting bias in an achievement item unless the missing knowledge is irrelevant to what is being measured (Green, 1972). These kinds of bias might better be called instructional bias or learning bias; they should not be confused with item bias, which is the result of irrelevant aspects of test items.

What might such an irrelevant aspect be? Suppose that the issue of the size of 9 or its square root came up in a word problem. Suppose further that for some reason the reading skills of the blacks were not as advanced, on the average, as those of the whites. Now if blacks performed less well, and if it was found that the reason was that they could not read as well, and if further it was determined that solving word problems using square root was not part of the curriculum, we would find it reasonable to conclude that the word problem based on square root was a biased item. The irrelevant aspect that caused the bias would be the use of reading which was not part of the curriculum.

Since it is very difficult for educators or psychologists to prove directly that there are genetic causes for mean differences in performance, they are reluctant to ascribe the cause to genes unless other likely possibilities have been eliminated. A basic issue, then, is whether the mean differences in group performance can be eliminated by using items which maintain the content validity of the achievement test but which avoid irrelevancies which result in group differences in mean scores. If that were not a possibility, most of the motivation for studying item bias would disappear.

One might argue that inclusion of biased items in a test is not detrimental as long as any items biased in favor of one group are balanced by items favoring each other group. For example, suppose a new curriculum is introduced in a particular subject, but only some students in a system experience the new curriculum. One might design a final examination for all students consisting of a core of items which all should know plus some items favoring each of the curricula. Angoff (Note 1) and Coffman (1978) suggested such a procedure. The balancing of items would make sure that students who were taught by one of the curricula would not have an advantage. Neither group's mean score would be higher. Jensen (1980a, p. 445) finds this unacceptable, however, and argues that no biased items should be used regardless of whether their inclusion would eliminate group mean differences.

Writing biased items is not difficult. All one has to do is to use slang or vernacular specific to a particular geographic location or subject-matter field. It would be easy to bias items to favor those who knew how to repair motors, or sew, or cook, or program computers. What is hard to do is to write items that are biased so little or so subtly that reviewers do not detect bias in evaluating the items until statistical data alert them to its presence. The issue is not whether biased items can be created. The issue is whether those who try as hard as they can to avoid bias still allow enough to account for the differences in mean test scores. Sensitive test construction agencies now screen items for the possibility that irrelevant factors might cause some cultural groups to be favored or handicapped. So one does not expect to see in modern mental tests flagrantly biased items. On the other hand, statistical tests of bias using enough examinees can detect such small amounts of deviance from equality that the deviance cannot be rationally explained. No irrelevant biasing factor can be discerned, as in the case of the square root of 9 item. What is needed then, is a trustworthy procedure for locating the bias that eludes the screening during item writing and editing but which, when noticed by means of statistical analysis, is readily discernible and easily rationalized with agreement among experts as to the source of the bias.

If biased items can be located that escape the scrutiny of item writers, editors, and reviewers, it may be that those items can be classified. Continued inclusion of items with biasing characteristics can then be avoided. Note, however, that every move in that direction makes the task of detecting bias that much more difficult. Even more subtle kinds of bias must be searched for in future tests.

Problems in Detecting Biased Items

A number of problems arise as we try to locate biased items. One is that judges do not discern accurately which items will be biased (Jensen, 1980a, pp. 371,554). Classic cases are the item in the WISC-R concerning fighting and the item in the Florida Functional Literacy Test (SSAT II) concerning filling out a check. Before data were examined, it was stridently proclaimed that these items were biased in favor of the white majority students. Later data analyses, however, showed that there was little difference between the performances of blacks and whites, and in some instances that the difference favored the blacks. As we shall see, it is typically the case that when item statistics suggest that items might be biased--they are discrepant--expert judges cannot explain the source of the discrepancy or agree on their explanations of it.

Another problem is that one must always keep in mind the matter of content validity. If an item, or a kind of item, appears to be biased but is relevant to what is to be learned or measured, then one cannot simply delete the item because one cultural group does better or worse than another on that item (Coffman, 1961). This should pose a problem for those (for example, Jensen, 1980, Rudner, Note 3, and Scheuneman, Note 4) who stress the statistical definitions of item bias. When they argue that any item which shows differences between students of equal ability should be removed, they may be destroying content validity. Complicating this issue even further is the fact that some students deliberately take courses designed to give them an advantage in certain kinds of activities. For example, many of my generation learned Latin because it was supposed to improve our comprehension and use of English. Assume that this was, and is, true. To suggest that all words with Latin roots be removed from test items to avoid biasing items toward those who have studied Latin would not be sensible. Similarly, one may feel uncomfortable about removing from a vocabulary test the word 'obligation' which research indicates is biased in favor of Hispanics. Content validity is a problem not to be treated lightly.

Which cultural groups should be studied for bias is another concern. Many different cultural groups are candidates for inclusion. Males and females are perhaps the most obvious, but differences between whites and blacks, and whites and Hispanics arouse great interest currently. Some studies have also considered such groups as Orientals, American Indians, Puerto Ricans, Cubans, Mexicans and people of differing socioeconomic levels. Urban and rural, north, south, east, and west could certainly be considered, and one must eventually realize that the number of possible cultural groups is infinite. Angoff and Ford (1973) even found discrepant items when comparing blacks from two different urban areas in the State of Georgia. Discrepant items have been found in contrasts of normal and handicapped (Rudner, 1978). No one seems to have treated in the literature which groups are most important to provide for or how many are optimum. Imbedded within the issue of which groups to consider is a problem of sample size. Some groups, e. g., American Indians, are too small for the use of some techniques even if the entire population is examined. Only rarely are biases between different minority groups studied. Some items may be biased in favor of Puerto Rican Americans as opposed to Cuban Americans, or Chinese Americans vs. Japanese Americans. If the phenomenon of bias, per se, were the issue, then logically these comparisons are as interesting as the comparisons with whites. That they are routinely ignored reflects the social rather than logical basis for the concern with item bias.

By definition, biased items are unusual items. They favor one group over another unfairly. However, some experts in the study of item bias argue that biased items are not just a little 'more so' than other items, but that they must be distinctly different. In a statistical distribution, they must be outliers (Coffman, 1978; Sinnott, Note 5). A gap should appear between their characteristics and the characteristics of other items (Shepard, et al., 1981). This reflects another problem in studying item bias. Defining just how extreme the difference between a biased item and its nonbiased counterparts must be for one to feel assured that the item is truly biased has been difficult. Some investigators would argue that this is a false issue, that bias is a matter of degree rather than a dichotomy (Lord, 1980; Rudner, Getson, and Knight, 1980a,b; Jensen, 1980a, p.575). Bias as a continuum and bias as a dichotomy lead to different practical implications. If bias is a dichotomy--some items are biased but others are not--then it might make sense to remove biased items (as long as content validity is preserved) if doing so eliminated mean differences between groups. However, if bias is a matter of degree--all items are biased but some more so than others--the practical problem is one of how much bias is too much. Statistical tests of significance are meaningless. The reductio ad absurdum is that mental testing cannot ever be done in a nonprejudicial manner. But that is ridiculous because we have a long history of useful and effective mental testing. So, if bias is a matter of degree, it becomes a matter of human judgment how much bias, in what directions, in how many items, is tolerable. Issues then are who should make the judgments, for what purposes, and how will we tell whether the judgments are valid. If group differences simply disappeared when the judgments were sound, to decide if they were sound would be simple. However, the bulk of the evidence suggests that group differences will not disappear in most cases if content validity is preserved. So it may be very difficult or impossible to get judges with vested interests to agree that the bias remaining in a set of items is satisfactory to all concerned.

These are some of the general and pervasive issues in studying item bias. There are many other concerns which are more specific to particular techniques. Some statistical issues present themselves in a wide variety of techniques. For instance, most analyses which compare item difficulty values between groups assume that the items come from a test which measures only a single dimension. It seems clear that if several dimensions are measured in a single test, one dimension may favor one group over another, and the items which measure that dimension would appear to be easier for the favored group. A major perplexity with this assumption is that agreement is lacking on a procedure to verify that the items in a test are unidimensional, or at least sufficiently homogeneous that multidimensionality will not be misinterpreted as bias in items. Another statistical issue is that of statistical significance vs. power. Some investigators have very large samples available for some of their groups--samples so large that trivial differences can be detected as statistically significant using certain techniques. Nearly half of the items appear to be biased in their analyses, but that does not fit the meaning of biased items' being outliers. (See, for example, Lord, 1977a.) On the other hand, some investigators use very liberal statistical significance levels, e. g., alpha = .3 (Scheuneman, Note 6) but with relatively small samples. Some (Shepard, et al., 1981) argue that significance level is not that important. For them, the distinctive gap between the outlying biased items and the rest of the items is the important issue.

Since in modern carefully built tests items have already been screened for glaring bias, detection of more subtle bias depends on statistical analysis of empirical data collected from examinees who attempt the items in realistic settings. The analysis of the data from those examinees takes many forms. Roughly fifty different methods, or variations on methods, have appeared in the literature recently. These methods are used to detect the items whose data are discrepant--the items which might be biased. Some investigators would label discrepant items biased with no further consideration (Groome & Groome, Note 7), but most active investigators at this time seem to regard the statistical analyses as only generators of flags which require verification before one concludes that the cause of the discrepancy is bias in the items. For example, Linn and Harnisch (1981, p. 116) say,

"(The presence of discrepant items) might be labeled bias. But to note that questions about the metric system are 'biased' against black students, and that story problems involving money are 'biased' in their favor is not very helpful. This observation implies the items are at fault. But it is at least as plausible that the model is at fault and/or that the 'bias' is due to instructional differences."

Others (Marascuilo & Slaughter, 1981; Breland et al., Note 2; Scheuneman, Note 8; Angoff, Note 9; Pine & Weiss, Note 10;) make a strong point of the fact that after items which are discrepant are detected by statistical or other routines, it is necessary for experts to examine identified items and determine what aspect of each of them is producing the bias. If it is not possible to determine what is causing the bias, these investigators are reluctant to call an item biased.

The bias-search techniques rather often call attention to items that behave as though they are biased, but examination yields no clue as to how they could be biased. This happens even when a variety of bias-search techniques all point to the same item as discrepant as occasionally happens. For example, Nungester (Note 11), using three different methods, found an item biased against females which involved determining the number of gallons of orange juice remaining in a tank if a certain percent were lost by leakage. He also found a problem involving multiplication of decimal fractions which was biased against males. It also happens sometimes when one technique picks a set of similar items as discrepant. Scheuneman (Note 6), using her C2 technique, found that items using negative language structures were biased. These 'biases' are not easy to comprehend. What introduces bias but is irrelevant to content is indiscernible. Willingham (1976), in describing the experience of the

Graduate Record Exams Board's efforts to isolate biased items in the GRE's, says, "That work has not been fruitful in discovering significant numbers of such items or in developing effective ways of identifying an unfair item prior to its use. . . In most cases it seems doubtful that these differences can be termed bias in any reasonable sense of the word. . ."(p. 18). We will see many more such examples in our review of the studies which attempt to apply bias-search techniques in practical situations.

The following review of the literature on item bias will describe a number of methods that have been developed to assist in detecting biased items. Relevant advantages and disadvantages of each method that have been noted in the literature will be presented. Following that, the studies that have attempted to apply the methods will be presented, including empirical tryouts of individual methods, empirical comparisons of methods using computer simulated data, empirical comparisons of methods using live data, and applications of bias detection methods in practical situations.

The literature leads the writer to the following important conclusions which seem consistent with the findings, but do not necessarily include all of their ramifications. First, judgment by people versed in detecting bias in items can be relied on to remove severely biased items. However, such judgments may err by removing items that are not biased, thereby removing from the available pool items (perhaps many) which might be perfectly useful. Second, items which are not severely biased require for their detection both statistical methods and human judgments. The statistical methods reveal that the items do not function in the same way for different cultural groups; the human judgments are necessary to ascertain that an agreed-upon irrelevant factor is the cause of the discrepancy. If no irrelevant factor causing the discrepancy can be agreed on, then the item cannot be classed as biased. In fact, it may be functioning very effectively in describing a relevant group difference. (Unfortunately, the statistical methods are so lacking in agreement with each other that no test can be cleared of the charge of containing biased items unless every imaginable statistical analysis for bias has been tried.) Third, even with the removal and replacement of items which are detected as biased by use of statistical and judgmental methods in concert, the differences in means between most cultural groups will not be eliminated. For a few pairs of groups, such as males and females or those taking different courses teaching the same subject matter, differences might be removed or "balanced" through careful use of items found to be biased. Fourth, the most beneficial use of statistical analysis may be to save items that appear superficially to be biased but actually yield the same results from the groups being compared. Fifth, when statistical analysis indicates that items are potentially biased, judgment reveals the irrelevant factor, and whole groups of items fall into the same category of being biased by the same factor, we have a case of some items measuring a different dimension from the others. In such a case, we should report scores on both dimensions if both are of interest. When only one item of a kind is found to be biased using both statistical and judgmental analyses, that item should be replaced by another item measuring the same content but unimpaired by the biasing element.

Discrepant-Item Detection Methods

The many methods that are used in searching for discrepant items can be grouped into nine general types of techniques for ease of exposition: (1)methods which involve comparisons of item difficulties, usually by plotting them on a graph, often after transforming the difficulties to a different metric (These procedures will be called transformed item difficulty or TID methods.); (2) analysis of variance (ANOVA) methods; (3) chi-square methods; (4) foil methods; (5) correlational methods; (6) item response theory methods (often called item characteristic curve methods and abbreviated ICC); (7) factor analysis methods; (8) methods based on experimental manipulations; and (9) methods of constructing tests to be sure they are unbiased. In the remainder of this paper, we will describe each of the methods and note the advantages and disadvantages of each as they have been described in the literature. Following the descriptions of the methods, we will consider studies which have employed them in order to compare methods or for the purpose of studying the items in real tests. Finally, we will attempt to summarize the findings of all of these studies and suggest practical implications and directions for further research.

TID Methods

The methods which involve plotting the item difficulty values of two different groups and comparing the difficulty levels have many variations. Many users of the methods transform the item p-values to equivalent normal deviates with a mean of 13 and standard deviation of 4. It is highly likely that this procedure became traditional simply because that was the internal procedure used by Educational Testing Service (ETS) for their analyses of test items. The people who first used this technique, Angoff and his colleagues, were employees of ETS and used PSAT items for their studies. Angoff used these transformed item difficulties because they had desirable characteristics, and others have merely followed suit. At any rate, the widely used designator for methods based on plotting of item difficulty values is Transformed Item Difficulty (TID) methods, and we will use that label here.

The complete method is as follows. First the set of items is checked for unidimensionality. If it is unidimensional, then the groups to be compared are equated on ability, usually by dividing the score range of the variable or a related measure into intervals and then selecting at random from the group with the larger number in each interval a number equal to the smaller group. For some pairs of groups, the ability distributions may be sufficiently similar that no further equating is necessary. Using the equated groups, the item difficulty (p value) for each item is calculated for each cultural group. The p values (more often, $1-p$ values) for all items and both subgroups are then transformed into normal deviates (z

values), before they are transformed into delta values by the linear transformation: delta = 4$\underline{z}$ + 13. A bivariate plot of pairs of item deltas is made for a test. The delta_x values represent the transformed $\underline{p}$ values for any one subgroup (e. g., males), and the delta_y values represent the other subgroup (e. g., females). This scatter plot of points normally appears in the form of an ellipse. Discrepant items are defined as those items or points on this graph which deviate from the major axis of the ellipse by some prespecified distance which, by definition, identifies only a small proportion of the items--those separated from the other points by a gap so that they are recognized as outliers. These items are potentially biased. The remaining items may be called concordant, as opposed to discrepant, thus avoiding prejudgment of the issue of item bias.

If the groups have been equated on ability, the items should fall along a 45 degree line through the origin from lower left to upper right. If the groups have not been equated for ability, or not fully equated--and some studies using the technique have not equated groups before calculating item difficulty values--then the ellipse will be displaced from the 45 degree line through the origin. Angoff and Ford (1973) provide the equation for the major axis of the ellipse and for the perpendicular distance of each point from the major axis. Some authors standardize these distance (D) values and report them in tabular form (Nungester, Note 11; Angoff & Stern, Note 12; Strassberg-Rosenberg & Donlon, Note 13; Hicks, Donlon, & Wallmark, Note 14).

Echternacht (1974) proposed a Kolmogorov-Smirnov type goodness-of-fit procedure to set confidence bounds around the major axis of the ellipse.

Items falling outside the confidence bounds are considered to be discrepant. Echternacht's study is the only known study of item bias using this technique, and in it he found no discrepant items comparing males and females.

An example of the result of plotting the points without equating the groups appears in Figure 1. The points are displaced from the 45 degree line through the origin and indicate that the group on the horizontal axis in general has less difficulty with the items. The ellipse is displaced to the lower right of the plot since the group plotted on the horizontal axis typically has higher delta (easiness or proportion correct) values. Items which are distant from the ellipse--have a gap between them and the ellipse--are the outliers or discrepant items which are potentially biased. Those distant below the ellipse favor the group on the horizontal axis; those distant above the ellipse favor the group on the vertical axis. Note that the major axis of an ellipse such as this is not the regression line through it. In fact, there are two regression lines but only one major axis. Also note that the plot of dots appears to be curved rather than a straight line. It can be argued (Lord, 1977) that if the groups are not equal in ability, logically the plot must be curved because everyone would get very easy items correct, and both groups would have the same delta values. Everyone would get very difficult items wrong, and both groups would have the same delta values for such items. Since one group is less able than the other, in the middle of the plot the points cannot fall on the line, but at the ends they must. Therefore, the line must be curved.

Insert Figure 1 about here

Green and Draper (1972) suggested another modification of the TID technique; it incorporates the item-test biserial correlations in order to adjust item difficulties for particular groups in a manner analogous to the way in which adjusted means are formed in analysis of covariance. The adjusted item difficulties are then plotted, and their distance from the major axis of the ellipse is taken as a measure of discrepancy. Hunter (Note 15) suggests obtaining an estimate of pure item difficulty by dividing item difficulty by item-test biserial r. Angoff (Note 16), with some reservation, seems to find merit in this suggestion. This modification of the TID procedure rests upon the contention that, given two items with equal differences in difficulty for two groups, the item which is more strongly related to the test score is more likely to be reflecting group differences in the behaviors the test assesses than is the item which is less strongly related to the test score.

Rudner and Convey (Note 18) attempted to equate on ability by simply standardizing delta values within the groups and then taking deviations from a 45 degree line. They plotted I_1 vs. I_2, where I_1 = (Delta 1 - Mean Delta 1)/S.D. Delta 1, and I_2 = (Delta 2 - Mean Delta 2)/S.D. Delta 2. In their data this procedure appeared to result in an abnormally large number of deviant items. Using the largest deviation from a 45 degree line in two groups chosen at random from a common population as the minimum deviation for a discrepant item, 25 out of 48 items are discrepant, 52%, when they compare hearing-impaired with normal examinees. Shepard, Camilli, and Averill (1981) indicate that the Rudner-Convey method does not linearize the relationship between item z scores for the two groups. Thus, curvilinearity between z values would be mistaken for bias.

Critical evaluations of TID methods. The TID methods have been used widely, and they have been criticized. Wright, Mead, and Draba (Note 18) point out that if the item p values are in the same order but differently dispersed for various cultural groups, this kind of discrepancy will not be detected by the TID method.

Lord (1977a), in a lengthy discussion of defects of the TID approach that justify an item characteristic curve approach points out the following problems with the TID approach:

a. The line of best fit to the scatterplot should not be a straight line, as we noted earlier.

b. The inverse normal transformation will not straighten out this curve if there is guessing.

c. The arcsin transformation will not straighten out the curve.

d. Items with different discrimination values will fall along different curves in a TID plot, so using only one curve for all items is inappropriate unless they are equal in discrimination.

e. Basically, p (or TID) does not reflect only difficulty; it also reflects group ability.

f. We cannot easily match our groups on ability in order to improve the analysis because we cannot soundly match on the score of the test whose items are being studied. We cannot match on a parallel form because we should be matching on true scores, not fallible scores, in order to avoid regression effects.

Hunter (Note 15, page 13) points out that the TID approach makes unwarranted assumptions. It assumes that the ICC's of the items are normal ogives with no guessing, that there are normal distributions of achievement in the various groups being compared, and that the items are equal in quality or discrimination. He states (pages 33-34):

> First, if a test has not been shown to be unidimensional, then there is no method based on the comparison of the statistics for a single item that is valid for the detection of bias in achievement items. Second, in particular, this means that not one iota of the evidence cited by Green or by Anghoff [sic] is valid evidence of bias against blacks in existing achievement tests.

Unidimensionality means that the score on a set of items measures only one characteristic. The difficulty with scores that measure more than one trait is that true differences between groups on one of the traits will make the items appear to be biased when on either trait by itself they would not appear to be biased. One way to evaluate unidimensionality is by means of factor analysis. Shepard, Camilli, and Averill (1980) accept a criterion recommended by Reckase (1979). He suggests that if the first unrotated principal factor accounts for 20% or more of the variance of a test's scores, the item parameters from an ICC analysis will be stable and the ability estimates will be accurate.

Rudner, Getson, and Knight (1980a) claim that while Hunter is correct about the assumptions, their empirical data suggest that the severity of these problems is not critical.

Shepard, Camilli, and Averill (1981) claim that whenever groups are not equal on the trait being measured, artifactual differences in item discrimination and differential guessing styles will appear to be evidence of item bias or will obscure actual item bias. Floor or ceiling effects could thus create false indications of discrepant items.

Finally, Rock and Werts (Note 19) claim that differences in p values could occur even though the covariance structure is the same across groups. In such cases, the difference in p values would reflect differential rates of development, not bias in items.

A virtue of the TID methods is that they are fairly easy to explain to laymen and help clarify the concept that item discrepancies are not mere differences in difficulty level of items, but are differences when groups have been equated on ability and differences from the general run of variability of difficulty levels that is found in a particular subject matter. A drawback to this method is that unless the techniques developed by Sinnott (Note 5) are used, only two groups can be compared at a time. If a number of different groups are to be compared, separate plots of all pairs of interest must be prepared. The principal advantages of the TID method are its simplicity and its low cost.

Analysis of variance methods

Many psychometricians have suggested analysis of variance (ANOVA) designs to determine the validity of a test for different groups (Cleary & Hilton, 1968; Plake & Hoover, 1977; Cardall & Coffman, Note 20; Fishbein, Note 21). All of these methods consider a test biased if the item x group interaction effect is statistically significant.

Cardall and Coffman (Note 20) used a two-factor (items x groups) ANOVA design with repeated measures on items. Cleary and Hilton (1968) extended this to a three-factor design (race, items, and SES nested within race) with repeated measures on items to determine whether a significant item x race or item x SES within race interaction effect could be detected. Neither technique suggested a way to identify specific items as biased, and neither seems to be very useful for a large-scale test program that seeks to identify biased items on its tests.

Fishbein's (Note 21) ANOVA method defined an item as biased if the difference in performance on the item for any two groups was significantly different from the difference in overall performance on the test for those two groups.

Plake and Hoover (1977) suggested transforming within-group item difficulties by an arcsin transformation to satisfy the assumption of homogeneity of variance and proceeding by analyzing each item by means of the interaction contrast:

$$\Psi_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..}$$

where

Ψ_{ij} = interaction contrast for the i^{th} item and the j^{th} group

θ_{ij} = arcsin transformed item difficulty of the i^{th} item and j^{th} group

$\theta_{i.}$ = mean arcsin transformed item difficulty for the i^{th} item across r groups

$\theta_{.j}$ = mean arcsin transformed item difficulty for the j^{th} group across the q items

$\theta_{..}$ = overall mean.

Even though many procedures are available to test all possible contrasts for atypical items, they recommend a Bonferroni-type comparison method, which is not as conservative as some of the other available techniques.

Critical evaluations of ANOVA methods. Critics have noted that the ANOVA approach to test item bias in which one searches for item x group interactions is troubled by the heterogeneity of variance of item difficulty values. The arcsin transformation used to reduce this problem (Cleary & Hilton, 1968; Plake & Hoover, 1977) has no ready interpretation from a measurement perspective, according to Wright, Mead, and Draba (Note 18). Angoff (Note 16) points out that the arcsin transform results in all the item-difficulty values having the same standard error (1/N) which makes it useful in ANOVA. However, its use in plots of difficulty values, is adversely affected by the fact that the arcsin values are bounded at both ends, yielding curvilinear plots.

Hunter (Note 15) showed that items of different difficulty will show an item x group interaction whenever the groups differ in achievement level even though there is no bias in the items whatsoever. Following Hunter's argument, consider Figure 2. It shows the ICC's for three items, A, B, and

---

Insert Fig 2 here

C, of varying difficulty levels but equal discrimination. The inverted normal distributions below the abscissa depict two groups of differing ability levels. The probabilities of correct responses for persons at the mean of each group's distribution of ability are indicated by dots on the curves. The vertical intervals, I, J, and K, show that the differences in the probability of getting correct responses (item p value differences) vary depending on the items' difficulties in relation to the groups' abilities. Items that are hard for both groups have small differences in p values (item C and difference K); items that are easy for both groups also have small differences in p values (item A and difference I); items whose difficulty is exactly centered between the mean abilities of the two groups have the largest differences in p values (item B and difference J). Thus, with items containing no bias at all and of equal discrimination value, some items may have larger differences in p values than others and therefore an appearance of an item x group interaction.

Rudner (Note 22) cites Jensen as showing that the ANOVA approach may be more sensitive to differences in ability than to cultural variations. That is, the significant ANOVA results from a comparison between races may be due to differences in mental maturity. Jensen matched subjects from different cultures on their mental age. After the matching, great reductions in the item x race interaction of the ANOVA were found. He also used ANOVA with two groups of whites whose score distributions closely matched those of the blacks and whites in his comparison between races. The results of this "psuedo-race" comparison were quite similar to the true cross-race results. Rudner (Note 22) also points out two more concerns with the ANOVA approach; one is a widely-recognized issue associated with making specific comparisons after a general ANOVA comparison is found to be significant. Since the specific comparisons were not hypothesized before the experiment, and there may be many specific tests, the probability of rejecting the null hypothesis of no difference is increased.

(Statisticians would say that the alpha level has been inflated for the post hoc comparisons used to determine which items are biased.) Second, in common with most other approaches, the ANOVA approach assumes that the total scores are unbiased. Perhaps these difficulties with the ANOVA approach have caused interest in it to decline.

One approach that is based on ANOVA is not used to detect biased items but rather to determine how likely it is that debiasing a test will remove the differences between the average performance levels of the groups being compared, differences that are the basis for the current general interest in item bias in the first place. Jensen (1980) describes the GD/I ratio, a ratio between F ratios. GD/I stands for group difference-interaction ratio: the F ratio for main effect for race is divided by the F ratio of item x race interaction. The same ratio can be expressed as a ratio between ratios of sums of squares or of mean squares. In studies of bias in which an ANOVA is computed, if the item x race interaction is not significant, the conclusion is that there is no bias. However, if that interaction is significant, then the GD/I ratio indicates how much larger the group main effect is than the item x group interaction. As the ratio approaches zero, the group mean difference becomes small and the group difference might be due to item bias. However, when the GD/I ratio is large, greater than 2, then the mean difference between the groups cannot be attributed to an unfavorable balance of biased items and cannot be reduced appreciably by eliminating some items or adding new items selected at random from the same general pool of items. GD/I ratios for groups of different sex are generally small, and, indeed, some tests are designed with a balance of biased items so that the means for sexes are about equal.

The GD/I ratios for race, by contrast, are usually large, and results like those of Flaugher and Schrader (Note 23) occur in which elimination of items has no apparent effect on differences between races.

Chi-square methods

Two main forms of chi-square technique are used, with a number of variations. Green and Draper (1972) advocate a chi-square technique which divides each group into fifths and then compares the proportion answering the item correctly in the same fifths but different groups. Scheuneman (1980) points out that this technique compares people in different test score ranges unless the groups are equal in ability. Hunter (Note 15) indicates that this dividing into fifths within groups is done because there is little overlap at the ends of the distributions when groups differ in ability. However, when each distribution is divided into fifths, one should plot item $\underline{p}$ values related to the mean ability of the groups in similarly numbered fifths or else artifactual bias will appear due to the nonlinearity of ICC's.

Scheuneman (Notes 4 and 6) advocates what she initially called a chi-square technique, in which the distribution of one group is broken into fourths or fifths and the cutoff values for that group are used to divide the score distributions of the other groups. In her procedure, an unbiased item is defined as one in which the probability of a correct response for individuals in the same ability level is the same regardless of group membership. Her procedure applies only to correct options, not incorrect options. For this reason, it is not a chi-square procedure, and Scheuneman later (1981) renamed it the C2 procedure. Shepard, Camilli, and Averill (1981) indicated that Camilli found that C2 statistic computed from proportions *correct* across ability strata for different groups will lead to different results than the C2 statistic calculated using proportions *incorrect*. Camilli then recommended fitting both correct and incorrect categories in a true chi-square. Both the Scheuneman C2 and the Camilli chi-square can be treated as though all differences were of equal significance regardless of direction, but in the study conducted by Shepard, Camilli, and Averill, C2 and chi-square values were also calculated keeping the signs so that differences which favor one group at one level but the other group at a different level would cancel out.

Critical evaluations of chi-square methods. A number of problems with the chi-square approach have appeared. Rudner (Note 22) states that one difficulty is that a large number of significance tests are made which is likely to inflate alpha and indicate that unbiased items are biased. (This is exacerbated by the fact that Scheuneman often uses an alpha level of .30 instead of the usual .05 or .01.) Rudner also indicates that the method assumes total scores to be valid measures of ability, and it appears to be unduly sensitive to differences in the total score distributions of the examined samples. Baker (1981) shows that Scheuneman's C2 method can make an item biased or unbiased simply by adjusting the relative sizes of the two racial groups tested. Further, he points out that Scheuneman has not shown that her C2 statistic is distributed as chi-square, and therefore results based on it should be used with extreme caution. Scheuneman (1981) replied to Baker that chi-square itself would be sensitive to differences in group sizes. Her solution is to sample from the larger group to reduce its size. She is not concerned about the sampling distribution of the C2 statistic since believes that she is trying to establish the truth of the null hypothesis and statistical inferences may be inappropriate in such situations. Ironson (Note 24) made an extensive comparison between the C2 test and chi-square tests. She indicated that the C2 technique seems more useful for very easy items, but that in both the chi-square technique and the C2 technique, (a) further analysis is needed to determine the direction and source of bias, (b) difficult items are not handled well, (c) cutoffs are arbitrary, (d) the procedures are sensitive to the distribution of total test scores, and (e) the procedures are sensitive to differential sample size of the cultural groups.

Scheuneman (1980) cites the following advantages for the C2 approach:

a. Domains are often unidimensional so multidimensionality is not an issue.

b. The procedure does not require large samples.

c. The procedure is simple and quick.

d. The procedure allows comparison among more than two groups at once.

e. The procedure should work well for mastery tests.

Scheuneman (1979) does, however, recognize the following constraints on the method:

a. The p value for the item within each ability interval must be less than 1.0 for at least one group.

b. The expected frequency for an interval must be at least 5, and the obtained frequency should be larger.

c. A minimum of ten to twenty observed correct responses per cell is necessary to avoid spurious results.

d. The smallest cell frequency should be the same size at each ability level.

She acknowledges that the obtained chi-square value in her C2 method may not approximate the chi-square distribution, especially if the samples differ in size or the cell frequencies are all very large. The chi-square values may not be usable as significance tests. For "significance tests" based on her C2 technique, Scheuneman (1979, p. 147) suggests that an alpha level of .05 will be appropriate, but she tends to use an alpha of .30. It is obvious from her struggle with setting an alpha level in this study, in which the groups were large and grossly unequal in size (and perhaps ability), that there are serious perplexities in using the C2 technique in deciding what level of so-called chi-square value indicates bias. Finally, Scheuneman (1977) points out that in using her technique, it is generally the case that the number of intervals which can be formed varies according to the difficulty of the items, the length of the test, and the size of the sample. Hence, it is possible that these three factors, instead of item bias, could determine whether an item is identified as discrepant.

Shepard, Camilli, and Averill (1981), in addition to noting that "chi-square" on proportion correct does not agree with "chi-square" on proportion incorrect, point out that differences in means within the matched intervals cause spurious results due to regression and reiterate Baker's (1981) warning that that large differences in the sizes of the samples confound the results.

In a highly technical methodological paper, Marascuilo and Slaughter (1981) clarify some of the issues as they describe a number of variations on chi-square methods including a method based on log-linear analysis. They believe that the chi-square techniques, while only rough approximations to the IRT model, are useful because they are easier to apply by those with relatively little statistical training and easier to explain to lay audiences. Also relatively small samples can yield statistically significant differences. However, in using the chi-square approaches, one must pay careful attention to the statistical power and the alternatives to which the power is directed. One of the methods they offer is similar to Scheuneman's method but is based on pairwise comparisons of $\underline{p}$ values for each different score level. They recommend allowing for the number of hypotheses being tested by dividing the alpha level by the number of score levels at which $\underline{p}$ values are being compared. If there is reason to state the direction of bias ahead of time, directional tests can be performed.

A second method suggested by Marascuilo and Slaughter uses the full chi-square, involving both successful and unsuccessful subjects. The authors note that the contribution to chi-square from the unsuccessful subjects is generally greater than the contribution from the successful, so the Scheuneman C2 procedure is likely to lead to an excess of Type II errors. This full chi-square test is a test against the hypothesis that at least one pair comparison is not equal to zero. However, it does not lead to ready interpretations of results.

A third method, attributed to Cochran, tests whether the same difference in proportions occurs in all paired comparisons, that is, that bias is uniform across groups with different score levels.

A fourth method tests for score group x ethnic group interaction. It examines whether the differences in p values for groups at various score levels are equal by comparing differences in p values to differences associated with least squares regression lines of probability of success on test score determined for each ethnic group. If a significant difference is found, post hoc hypotheses can be tested on the individual contrasts between ethnic groups at specific score levels to display the nature of the interactions. According to Marascuilo and Slaughter, a significant result from the overall test indicates that the discrimination value of the item differs for the ethnic groups, even though the difficulty level may be the same.

Marascuilo and Slaughter's fifth method is a log-linear analysis of ethnicity interaction. It is assumed that the probability of answering correctly is an exponential function of the sum of four parameters. Taking the natural log of the probability results in the function being linear in the sum of those parameters. Analysis is accomplished by using one of several computer programs which culminate in a statistic which is compared with the chi-square distribution. A significant result indicates that item discrimination differs between the ethnic groups. Marascuilo and Slaughter point out that their methods four and five can be used interchangeably, but method four is easier to follow and would be preferred in use with lay audiences. The two methods appear to be equal in power.

These authors illustrate a sixth model which focuses on the quantitative nature of the score-grouping variable. It is a procedure of stepwise regression in binomial variables and ordered categories. The method reduces a total chi-square to its minimum by determining how much variation is explained by models that become more and more complex as the technique proceeds. The following hypotheses are tested in order (Marascuilo & Slaughter, 1981, p. 243):

a. Homogeneity of proportions.
b. Regression lines possess a linear component.
c. Regression lines are parallel in their linear components.
d. Regression lines are not parallel in their linear components.
e. Regression lines possess a quadratic component, etc.

When chi-square becomes nonsignificant, the process is ended. The analysis is usually done on a computer. Prototype programs are available from Marascuilo and Slaughter.

Marascuilo and Slaughter recommend that more than one of the above six methods be used as appropriate. All of them are lengthy and benefit from computer use. However, their first, third, and fourth methods should suffice for most investigations.

Foil comparison methods

Veale and Foreman (Notes 25, 26) suggested examining the differential response patterns of various groups of individuals to the item foils in order to find alternatives which overly attract or repel a particular group. Jensen (1976, p. 342) supports this type of thinking when he states that "cultural bias leads to the expectation that whites and blacks should make different errors among the multiple-choice distracters of the items they get wrong." Veale and Foreman suggest applying a chi-square statistic to detect differential-response patterns. This procedure is most clearly described in the computer-program documentation for the process (Kohr & Tung, Note 27). Setting up a bivariate table with cultural groups (rows) and foils (columns) is the initial step; it enables one to test whether the proportional response to item foils is equal across cultural groups. Complexities can be encountered with the chi-square statistic when the expected frequencies in some of the cells within the groups x foils matrix become small for any one item (i. e., any one of the expected cell values is smaller than five). Otherwise, one may proceed with the conventional chi-square test. Further, they recommend under the same sample size limitations that a Cramer's V statistic (which uses the chi-square statistic) be employed as a measure of association (or homogeneity of response distributions) across groups for an item. If the response frequencies in 20 percent or more of the cells are below 5 or any one cell is zero, then several statistics developed by Goodman and Kruskal are recommended in the Kohr and Tung (Note 27) paper to do basically the same thing as the chi-square method. In both the chi-square and Goodman-Kruskal techniques, if the null hypothesis is rejected, then the item is classified as biased; if it it not rejected, then it is unbiased. Veale (Note 28, p. 1) indicates that this method "is based on the idea that students' responses to the wrong answers (the 'foils') can provide more information concerning the cultural validity of an item than their responses to the correct answer (which confounds bias with ability or knowledge of the objective)." Inherent in these arguments is the assumption that all subgroups of a population should be equally (proportionally speaking) attracted to all foils. Veale and Foreman suggest that for each foil on each item, "pull indices" be calculated. The pull index for a foil for a particular group is the proportion of that group responding to each foil out of the total number of foil responses.

Critical evaluations of foil comparison methods. The foil response method has been critically evaluated though it seems not to have been used. Rudner (Note 22) pointed out some of its advantages. The method does not use total test scores so there is no assumption that the test, itself, is unbiased. Rudner points out, but without explicit support for the assertion, that it is the only method available for criterion-referenced mastery tests. Rudner also points out that the method is one of the few that pinpoints the source of bias to a specific distracter and that this method makes no assumption about the accuracy of the item $\underline{p}$ values or the total scores.

The method has many difficulties, however. Rudner (Note 22) pointed out that it is not sensitive to bias in the stem, the directions, or the subject matter of an item, and that there are so many significance tests that the alpha level is inflated. The result is that unbiased distracters may appear to be significantly biased. Petersen (Note 29) claims that differential response to the distracters may be a function of ability, not just bias. Scheuneman (1980) raised the same issue, explaining that the most able students will be drawn to the distracter which is closest to being correct causing it to appear to be biased and other distracters to be biased in the opposite direction. She (1980) also claimed that the method assumes that the groups are of equal ability. While Veale (Note 30) claims at one point that this is not so, on another page of the same paper he states that the procedures are clearly subject to the disturbing effects of ability differences between groups.

Veale, himself, (Note 30) criticizes the method as he lays the groundwork for a new method. He points out the following concerns:

a. Biased foils will not be detected if all are biased.

b. If one or two are biased, they will be detected only if (1) the groups are of equal ability or (2) the foils are equally attractive.

c. If different foils are biased against different groups, only the most seriously biased will be detected and then only if groups are equal in ability.

d. The model is seriously affected by ability differences between groups.

e. The model is seriously affected by differences in the attractiveness of the various foils, i. e., foil difficulty.

f. The accuracy of estimate of "pull indices" is low for easy items.

g. The accuracy of estimate of "pull indices" is not good for small amounts of bias unless the groups are of equal ability.

h. Sample sizes of about 250 are needed in each cultural group, so the method cannot be used on small samples.

Nowhere in the discussion of the method in Veale's 1978 paper (Note 30) is the matter of studying several different ethnic or other groups at once given full consideration. It must raise interesting problems of some distracters' being attractive specifically to particular groups but not to others. Those issues deserve careful discussion. To circumvent some of these difficulties, Veale developed a new method, the foil response curve method.

In this method, a curve is developed for each distracter showing its pull index for persons of various ability levels, and these curves are compared for different cultural groups. If the response curves differ, it suggests bias. Either the total test score or an outside criterion is used to sort people into various ability levels. The curves are compared by obtaining regression coefficients for fitting straight or curved lines to the pull indices for various levels. Then a $\underline{t}$ test is used to test the difference between regression coefficients. The practical significance of the bias is evaluated by calculating the area between the foil curves and summing over the foils. No reports have been found of the use of these foil response methods on real or simulated data except for examples of their performance on individual items used to illustrate the methods.

The assumption that all subgroups of a population should be equally (proportionally speaking) attracted to all foils is troublesome and seems contrary to accepted test-construction practices. Good item writing requires the item writer to think of various groups of examinees and to write foils that are attractive to members of one group or another who have not learned the correct response. For example, good test construction will often use two devices to develop sound foils. One is to give an open-ended or completion item in order to see what kinds of errors are commonly made. These errors then become distracters in the multiple-choice test format. The other is to think of the reasons for the kinds of mistakes that various sorts of examinees might make. For example, the student who relies on incorrect information which he brought with him to the instruction instead of relying on what he has been taught will probably make a mistake. That incorrect information becomes the basis for a distracter in order to identify students who have not learned to avoid it. A case in point would be the English teacher who uses split infinitives in foils to be sure that her students have learned that the correct form does not split the infinitive. Another would be the English teacher who uses street or ghetto dialect in her foils to be sure that the students have learned that those responses are incorrect. The Veale procedure would call these kinds of items biased and would eliminate them, eliminate the foils containing split infinitives or ghetto language, or else would adjust the p level of the item to pretend that students were not making these mistakes. Any of these procedures is unsound, and this is a fundamental and perhaps insurmountable obstacle in routine use of this technique.

It seems that, used with great care, Veale's foil response methods may be helpful in identifying the cause of bias in items. But, the logic and the merit of correcting item *p* values and test scores for "overpull" are questionable and require repeated iterations. It is not clear from any of the literature available on this procedure exactly what the results of such a process would be.

Correlation methods

A variety of methods for using correlation coefficients to locate bias in tests has been suggested. The simplest, and probably least effective, is to compare reliabilities of a test when a reliability coefficient is estimated separately for each cultural group. Green and Draper (1972) seem to be the first to have suggested that tests with different reliabilities for different groups exhibit bias. They argued that a test which was less reliable for one group would yield less accurate scores for that group and, thus, in this particular sense, be biased. They propose to locate biased items by estimating the contribution of each item to the size of the Kuder-Richardson formula 20 reliability of a test, given the remaining items, and apparently they would contrast these contributions between cultural groups, although they did not fully explain their approach.

Green and Draper pursue this approach very little, probably because their data led them to believe that this kind of bias is not very common. In the data they examined, which included seven different cultural groups, the reliabilities computed within groups ranged from .90 to .93. A range as narrow as this clearly does not offer sufficient opportunity for systematic variation between groups. The ineffectiveness of this method is corroborated by Jensen (1976) who chose two tests for study, one he felt would generally be accepted by experts as relatively free from bias, and one he felt most likely would be biased. The test that was chosen for its susceptibility to bias yielded reliability coefficients of .96 in both black and white groups, while the relatively unbiased test yielded reliabilities for whites of .90 and blacks of .86. Results such as these do not convince one that tests will reveal their bias in reliability coefficients. However, it may still be true that tests which have highly different reliabilities for different cultural groups do so as a result of bias. Nonetheless, even if these techniques did produce results that suggested a test was biased against a particular group, they may not be effective in the detection of biased items.

Jensen (1976) suggested several techniques that use the correlation statistic. In one, he claimed that a test that measures different things in different groups, which is in accord with one basic definition of bias, will be likely to have different degrees of correlation between raw score and age in months for the different groups. To our knowledge, no use of this method other than Jensen's has been described in the literature.

Another correlational technique used by Jensen was that of estimating the difficulty level of each item of a test separately for each of two cultural groups, correlating the rank order of these p values, correcting the correlation for attenuation (using the reliability of the rank order of p's within each cultural group) and contrasting this correlation with the correlation within a single cultural group. He showed this method to be sensitive to cultural differences by considering data from the administration of a culturally sensitive test to school children in London, England, as opposed to age-matched white children in California. The rank orders of a number of items differed markedly in this contrast. Cardall and Coffman (Note 20) used this same procedure as a follow-up to ANOVA. They gathered several groups of examinees from urban New York City and the rural Midwest, as well as several locations in the Southeast in which only blacks were registered to take the College Board's SAT. The correlations of item difficulties within one of these groups were noticeably higher than the correlations of item difficulties between rural Midwesterners and blacks from the Southeast or between urban New Yorkers and blacks from the Southeast.

A third correlational technique used by Jensen was the correlation between differences in item difficulties between adjacent items in a test. Jensen argued that it would be remarkable, indeed, to find the rank orders of difficulty and of differences in difficulty of adjacent items highly similar if items differed in bias for the cultural groups being compared.

A fourth correlational technique, often called the item-discrimination technique, uses the item-test correlation coefficient as the indicator of bias. Item-test correlations are used in developing tests in order to increase test reliability. Tests that are supposed to be measuring a single trait should be composed of items each of which measures that trait. Thus, the items which correlate most highly with the total test score are those which seem to be best to include as measures of the desired trait. Indeed, if the items with the highest item-test correlations are included, the final total score will have a much higher reliability coefficient, i. e., it will measure the characteristic with much less measurement error.

Usually the item-test correlations are obtained by testing a representative sample of the people who are later to be measured by the test. Thus the items for a test for use by eighth graders would be analyzed during the process of test development by giving them to a representative sample of eighth graders and computing item-test correlations. Green and Draper (1972) suggested that different items might be chosen for a test if the item analysis were done on minority groups instead of majority groups. They suggested analyzing the items on each cultural group to which the test is to be applied and choosing for the final form the best half of the items from each cultural group. Those items which are chosen on the basis of one group but not some other group are regarded as biased. Hunter (Note 15) criticized this approach on the ground that for groups differing in ability different items will be selected because the p value of the item directly affects the item variance even when item intercorrelations are equal. Unbiased items thus are differentially chosen for the two groups and are regarded as biased.

ICC methods

A variety of methods of detecting item bias has been developed based on item response theory and item characteristic curves. The idea is that for items that measure the same single characteristic or latent trait, statistical parameters which describe the functioning of the item in measuring that latent trait can be determined. If the item measures only the one latent trait--is unidimensional--then, according to this theory, it does not matter on what group its parameters are estimated. They will be the same after equating regardless of the groups tested (Lord & Novick, 1968). If different groups yield different item parameters, then it is concluded that the item measures different things for different groups. In this sense, it measures group membership along with the latent trait it is supposed to measure. In brief, the same item parameters should be obtained regardless of the group if the trait is unidimensional. Different parameters indicate the presence of bias.

Three item parameters are of primary interest. One, usually called a, represents the effectiveness of the item in discriminating between persons who have a high level of the trait and those who have a low level of the trait being measured. The second, usually designated b, represents the level of trait necessary to have a medium probability of obtaining the correct response to the item. In essence, the b value represents the difficulty of the item. The third, usually designated c, is the proportion of examinees of minimum ability who answer the item correctly, presumably by guessing.

A variety of methods of detecting potential bias has been developed based on these item parameters. Hunter (Note 15) has pointed out that all of them are based on the assumption of unidimensionality. Lord (1977) points out that even if a test measures only a single trait, if something occurs corresponding to teaching certain items to one group but not to another while other items are taught to all groups, dimensionality is increased. If we ignore this fact and analyze all items as though they were unidimensional, we cannot expect all ICC's to be the same for the groups since the expectation that all groups would have the same ICC was based on the assumption of unidimensionality. Increased dimensionality due to differences in instruction or learning could be a severe challenge to the study of item bias if language differences, community differences, school differences, or cultural differences caused some content to be learned better by some groups than others. Any method requiring unidimensionality would be ineffective according to Hunter. Of course, it must be recognized that items which reveal that some groups have learned certain content that has not been as well learned by other groups are not "biased" items if the material is relevant to the school's objectives, as we noted earlier. "Biased" items are items on which groups differ in performance for irrelevant reasons _when they know the material equally well_. One cannot expect psychometric techniques to detect biased items when the items are not biased but the learning is. That bias is a matter of learning, not of item construction.

Hunter (Note 15) also noted that some authors, e. g., Green and Draper (1972), attempt to use the regression of item performance on test score as an approximation to the item characteristic curve of item response theory. Jensen (1980a) also suggests using the regression of item performance on test score. He proposes using a chi-square test of the null hypothesis of no difference between the item-test regressions of the two groups. He suggests that this method requires very large samples in both groups, and that it will be weakened if both groups' scores are not proportionately distributed in the same way within each score interval. Usually they will not be distributed in the same way if there is a marked difference between the overall groups' means. The overall lower group will tend to have the lower mean score within any score interval. The amount of bias can be estimated, according to Jensen, by comparing the two groups' means on the test when the biased items are present and when they are removed. According to Hunter, these item-test regression methods require the assumption of equal reliability of the scores in the groups being compared. Even with reliabilities in both groups in the .90s, differences in reliabilities between the groups can cause the appearance of bias where none exists. For example, reliabilities for the same test can differ between groups due to differences in variance between the groups, or for other reasons. Rudner (Note 3) criticizes Green and Draper's approach because it does not provide a satisfactory quantification of degree of bias. One looks at the curves and subjectively decides when a difference is large enough to cause concern.

One-parameter model methods. One approach to using ICC's to detect bias is to adopt a model of what is meant by "good measurement," to collect data from test items administered to different cultural groups, and to evaluate whether the data fit the model equally well in the different groups. Wright, Mead, and Draba (Note 18) recommend this procedure, under the assumption that the item a parameters and c parameters equal 1.0 and zero, respectively, a test-theory model developed by Rasch. They test the goodness of fit of the items to the model, and if the residuals due to poor fit are large, the item is regarded as biased. The advantages of this method are claimed to be:

a. The method divorces item difficulty from person ability.

b. It uses the only realistic transformation of item p values, the logistic transformation.

c. It provides useful asymptotic estimates of the variances of parameter estimates.

d. It makes possible the separation of biased items from items which do not fit the model for other reasons.

e. It makes possible the specification of the size of residuals when the fit is good.

Shepard (Note 31), however, points out that Wright, Mead, and Draba's residuals will identify lack of fit for reasons other than different ICCs as though the lack of fit were bias. Divgi (Note 32) criticizes the Wright, Mead, and Draba goodness-of-fit measure on two grounds. First, he cites Gustafson (Note 33) as having shown that there is no statistical test of item fit in the conditional maximum likelihood approach. Wright and his colleagues offer an unconditional maximum likelihood method for estimating person ability and item difficulty parameters. They also offer an approximation, manageable without a computer, which assumes normal distributions of ability and of item difficulties. If the condition of normal distributions is used, then it appears that it is not possible to test goodness of fit statistically. Second, the Wright, Mead, and Draba approach evaluates goodness of fit to a linear model, but the effect of guessing is best described by a quadratic function. Divgi (Note 32, page 4) states, "Guessing raises the ICC at all abilities, the increase being largest at low abilities. Its contribution may be described by a quadratic function." Divgi develops a test of goodness of fit based on a quadratic model and finds it to be much more powerful than the test used by Wright, Mead, and Draba. In an example based on data from the Metropolitan Survey Reading tests, Divgi's test finds 69% of the items fail to fit the model. The goodness-of-fit test recommended by Wright, Mead, and Draba finds only 16% fail to fit. It remains to be seen whether differences in residuals from a goodness-of-fit test such as Divgi's are useful in detecting item bias or whether the one-parameter ICC model fits enough operational test data well enough to be a useful model when rigorous tests of goodness of fit are employed.

Durovic (Note 34) introduced another goodness-of-fit method based on the one-parameter model. It is described in Scheuneman (1980). First, the goodness of fit to the one parameter model is estimated for each ethnic group using the procedures of Wright and Panchepakesan (1969). Then bias is defined as a difference in mean square fit greater than 1.0 between the groups. There are some concerns with Durovic's method, of course. Divgi's observations about the power of these goodness-of-fit tests apply. Also, some authors (Tinsley and Dawis, 1975) suggest that there should be at least 500 people in each cultural group.

Wright, Mead, and Draba (Note 18) suggest that item bias can be studied by obtaining good estimates of the difficulty levels of items from a sample and determining whether the same difficulty levels and discrimination effectiveness occur in another group of people. A chi-square test of fit of the item to the model for a particular group tests "how similar item i is to the typical item used to represent the variable of interest" (Note 18, page 15). The residuals may be studied with ANOVA to determine whether the lack of fit is due to differences in difficulty or discrimination. If the lack of fit is due to either, "clinical investigation of the implicated items in terms of their content and the characteristics of the sample of persons should help diagnose the source of bias. It might then be possible to discard the data from the defective items and extract fair measures without retesting" (Note 18, page 17).

Furthermore, Wright, Mead, and Draba claim that an extremely important feature of the method is that person fit can also be analyzed so that for each person a check can be made on the quality of the observation obtained from him on each item. Person residuals can also be analyzed, and items biased for a person can be deleted for that person. This introduces a whole new dimension to the analysis of bias in items. Now we are not looking for items biased against particular groups, such as minorities, people with certain kinds of handicaps, people from certain parts of the country, people of certain socioeconomic levels, people with certain dialects, and the like, but instead we are looking at the possibility of bias in every item for every individual. When items unfair to an individual are located, they are to be deleted from the test score, and new unbiased measures are to be developed for each individual.

To see where this method might lead, suppose we have an arithmetic test consisting of items that involve using the four fundamental operations on integers, decimals, and fractions. Now suppose Mary knows how to do problems as long as they do not involve fractions, Billy can do problems as long as they do not involve decimals, Sammy can't divide, and Virginia has not learned her multiplication tables. Using Wright, Mead, and Draba's person-fit method, each of these students will be tested only on the items that "fit" for him or her, and each will get a good score on arithmetic because he or she does well everything except the things he or she can't do at all. Will such scores have any useful meaning as measures of ability to do arithmetic?

Draba (Note 35) suggested using differences in the difficulty parameter of the Rasch model as the indicator of item bias. If the assumptions of unidimensionality, equality of item discrimination, and absence of guessing, which are demanded by the one-parameter model, are met, one would expect the Draba method to yield essentially the same results as the TID method using groups equated on ability.

Three-parameter model methods. Several different approaches are based on the three-parameter model. Lord (1977b) described briefly an unpublished approach used by him and Marco. They used ETS's LOGIST computer program to obtain item parameters. Then item characteristic curves were plotted for different cultural groups. Items were designated biased if one group performed better than another at every level of the trait, if the item discriminated well for one group but not for the other, or if the item discriminated noticeably better for one group than the other.

Pine and Weiss (Note 10) used a less complete version of this method. They defined bias as the difference in the parameters reflecting item difficulty when the parameters are based on different groups. They noticed in a computer simulation using this method that items with high discrimination parameters which were biased (had different difficulty parameters for two groups) produced greater differences in subgroup validity than did similarly biased items with lower discrimination values. This suggests that biased items that are not very discriminating will have little harmful effect in terms of differential validity for subgroups.

In another paper published in 1977, Lord provided greater detail on the use of LOGIST and the three-parameter model to detect biased items

(Lord, 1977a). The steps he suggested there are as follows:

1. Obtain data from equal numbers of examinees from two cultural groups, using at least several thousand cases in each group.
2. Run LOGIST on each group.
3. Equate the item parameters.
4. Test the significance of the differences in ICC's.
5. Eliminate items whose curves are significantly different.
6. Combine the two groups.
7. Run LOGIST on the combined data on the remaining unbiased items. This estimates the parameters of both groups on the same scale.
8. Divide the data into random halves, ignoring cultural groups.
9. Run LOGIST on the two halves using all of the items.
10. Compare ICC's with significance tests to see what value is an appropriate alpha level, since any differences between curves from these two groups should be due only to chance.
11. Run LOGIST on the separate cultural groups using the parameters of step 7 as given but including all the items.
12. Test differences between ICC's using the alpha level located in step 10. Any items reaching that level are biased.

Lord (1980) suggested yet another procedure for using the three-parameter model with LOGIST. The steps are as follows:

1. Estimate approximately the item parameters for all groups combined, standardizing on the b_i and not on theta.
2. Fixing the c_i at the values obtained in step 1, re-estimate a_i and b_i separately for each group, standardizing on the b_i.
3. For each item, compare across groups the item response functions or parameters obtained in step 2. (page 217)

Statistical tests of significance of differences between a_i and between b_i are provided, and a chi-square test is described for testing both hypotheses simultaneously. If many of the items are seriously biased, the test may not be unidimensional. A procedure for purifying the test which is much like the procedure described in the previous 12 steps is recommended.

A contribution from Rudner (Note 3) is to use the area between ICC's as the measure of bias. He approximated the integral of the difference between the curves, a procedure that put the difference on a ratio scale in a sense, and it included differences in difficulty and in discrimination. It measured degree of bias instead of just detecting which items are biased. The method has some drawbacks, of course. As with all three-parameter ICC models, it requires large numbers of subjects and items. It is not applicable to items with extreme $\underline{p}$ values or to items that do not fit the ICC model. The direction of the bias is not identified, and, in fact, the ICC's may cross. Also, large areas may be due to things other than item bias, such as local dependence, multidimensionality, or poor parameter estimates. Divgi (Note 36) pointed out that the area between the curves lacks intuitive meaning, and its value depends on the scale being used. If abilities are rescaled so that the ability difference between two given persons is doubled, the area between ICC's gets doubled also.

Perhaps one of the most troublesome disadvantages in practical use of the three-parameter ICC approach is the demands it makes on data and budgets. Scheuneman (1979) pointed out that it requires on the order of 2000 cases and 40 to 60 items per test. Carlton and Marco (Note 37) indicated that even at ETS, where to a large extent item response theory was developed, ICC methods are so expensive to apply that they will probably be used in connection with the SAT only for special studies. Even the one-parameter model faces some difficulty on this score. Scheuneman indicated that while Wright believes that 100 subjects are sufficient in each group, Tinsley and Dawis (1975), as we noted in the previous section, recommend 500 cases in each group when using the one-parameter model to be sure of parameter estimates when groups are as diverse as are often of interest in the study of item bias. The ICC methods are not regarded as generally appropriate, then, for short tests, small groups, or small budgets.

Linn and Harnisch (1981) attempted to circumvent some of these difficulties. In a 1980 study, Linn and his colleagues using the three-parameter model found that when they had samples of 2000 cases they obtained reasonably small standard errors, but when they had only around 500, the standard errors were so large as to be of limited use. The problem is exacerbated by the fact that it is often difficult to find groups of minority examinees as large as 2000. To permit using ICCs to study item bias with small numbers of examinees, Linn and Harnisch suggested that one first obtain item parameter estimates for the three-parameter logistic model using all available cases. From these, obtain the estimated probability that each person would answer each item correctly, given his ability estimate. These estimates of probability of being correct are compared with the actual observed rates of correct responses for the particular target group or for subgroups divided according to ability level. By comparing probability of being correct with observed frequency of correct responses at different ability levels, one can ascertain whether an item with a small observed difference overall may have large positive discrepancies for some ability levels but large negative discrepancies for other ability levels.

Linn and Harnisch note some difficulties with this method. The estimates of the item parameters will be influenced by the presence of the minority groups in the total sample of examinees. Also, the optimum grouping of ability levels to be used in studying discrepancies at various ability levels is unknown. The discrepancies that are found cannot be compared across testing situations. The most serious disadvantage, according to Linn and Harnisch, may be the fact that the probabilities are based on estimated item and person parameters rather than the actual parameters.

Factor analysis methods

Several different factor analysis methods have been suggested. An intergroup factor analytic approach to the examination of achievement tests for bias was suggested by Green and Draper (1972). The method appears to be related to Tucker's (1958) interbattery method. This advanced factor-analysis technique essentially compares factor-analysis results obtained from two separate groups of examinees by doing factor analyses on both of them simultaneously. The method yields some factors which the groups have in common, some factors (if they exist) that are operative in one group but not the other, and the specifics, communalities, and other paraphernalia of factor-analysis results. The aspect that should be different and useful about this type of factor analysis in studies of item bias is the possibility of each group's having some factors of its own that it does not share with other groups in answering these items. If this occurs, the test is regarded as biased. The items which have substantial factor loadings on the factors unique to one group are looked upon as biased items.

Green and Draper found that they could not do intergroup analyses on more than two groups at once due to computer programming problems. They reported several two-group analyses, compared results from this method with results from other methods, and reported several contradictions. Green's further work (Note 38) involved computer simulations which depended on item-test correlations, but eventually worked back into intergroup factor analyses. Green regards his results as highly tentative at this point (Note 38) so use of the intergroup method by anyone planning an application should be considered with caution.

Merz (1977, Notes 38, 39) suggested a factor analysis method which also uses ANOVA. The item intercorrelation matrix is computed for persons across groups. It is factor analyzed with rotation to simple structure. Factor scores are derived for each person. Then an ANOVA is done on factor scores with groups as the independent variables. If any factor has a significant main effect or is involved in a significant interaction, it is regarded as a biased factor. Any item loading heavily on a biased factor is regarded as a biased item. As far as can be found, this method has never been used.

Merz (Note 40) used another method involving factor analysis. In this method, group membership is entered into the item intercorrelation matrix as dummy variables. After factor extraction and rotation, factor loadings are examined for items which have large loadings on factors that also have large loadings on group membership.

Rock and Werts (Note 19) based an approach on the work of Joreskog on confirmatory factor analysis. It is now possible to test statistically for (a) goodness of fit of a priori models of factor structure, (b) equality of units, (c) accuracy with which each construct is measured, and (d) invariance of the above three concepts across populations. If they are not invariant across populations, the conclusion is that there is a form of psychometric bias. The actual items causing bias are not specified, but it may be possible to detect which kinds of items are biased if different parts of a test are composed of items of different kinds.

Another factor analysis method suggested by Jensen (1980a) evaluates only the pattern of loadings, not their size. For this approach, one first splits the major and minor groups into random halves. Then principal components analyses are done on each of the four sets of data. The correlations of loadings on the first principal component are calculated for all six combinations of groups. The two within-group correlations are then averaged using Fisher's $\underline{z}$. This gives the average within-group correlation. The other four correlations are averaged to get the average between-group $\underline{z}$. A chi-square test is then used to compare within- and between-group correlations. One disadvantage of the method is that it requires at least 200 cases in each group--i. e., a total sample of 800--and requires that the group sizes be equal.

Jensen (1980a) suggested another factor analysis method. It is a test of the overall difference between loadings on the first principal component of the major and minor groups: it tests both the pattern and size of the loadings. The principal-components factor analyses are first done on each group separately. The loadings are then converted to Fisher $\underline{z}$ coefficients. A chi-square test is used to test the hypothesis that the loadings do not differ between groups.

Still another method suggested by Jensen (1980b) is based on factor analysis of between- and within-family correlations between tests using pairs of siblings from different cultural groups. Scores on each test are first standardized by age within cultural groups. Then a check is made to be sure that the measures have interval scales. The between-family correlations among the scales are obtained by correlating scores of sibling pairs. The within-family correlations among scales are obtained by correlating the differences between scores for siblings on one scale and the differences between scores for siblings on another scale, with the younger siblings's score subtracted from the older in each case. Those correlations are corrected for unreliability, and then the between-family and within-family correlation matrices are factor analyzed separately by cultural group. The same general factor should be found in both groups and within- and between-family if there is no bias. A factor congruence coefficient is calculated to verify this.

Critical evaluations of factor-analytic methods. Rudner (Note 22) believes that the factor analytic methods are advantageous because they deal with underlying traits, the Green and Draper method partitions variance into culture-specific and culture-common sources, and the Merz method partials out variance caused by factors such as socioeconomic status, intelligence, and sex. Scheuneman (1980) claims that factor analysis methods circumvent the problem of differences in ability among samples. In fact, the factor analysis methods seek to determine whether the test(s) have the same meaning for the two groups, and in this sense these methods have the same intent as all the other methods.

Others are less enthusiastic about these methods. Hunter (Note 15) claims that the nonlinearity of ICC's invalidates the intergroup factor analysis proposed by Green and Draper. Scheuneman (1980) points out that factor analysis has difficulty in getting clear factor structure using item data. Shepard, Camilli, and Averill (1981) argue that in Merz's approach, true differences in group ability will appear as item bias. Rudner (Note 22), himself, recognizes some problems with the factor analysis methods. They require large numbers of subjects, a fact which often goes unrecognized, and there are a number of decisions to be made about the particulars of technique, e. g., the kind of correlation coefficient to use. The phi coefficient tends to produce artifactual difficulty factors unless all items have the same *p* values, and the tetrachoric coefficient tends to be unstable. Rudner, Getson, and Knight (1980b) also note all the decisions necessary in factor analysis as being a problem, and Rudner and Convey (Note 17) had what was regarded as an unsuccessful tryout of a factor analytic method.

Methods based on experimental manipulations

Several methods have been described, suggested, or hinted at which involve deliberate experimental manipulations of subjects rather than comparisons based on particular aspects of already-available data. Green (1975) suggested that bias in criterion-referenced tests might be detected during the process of studying the validity of such a test. One indicator of validity of a criterion-referenced test is its sensitivity to instruction. Such a test is usually tied closely to instructional objectives, and the items are supposed to be based on content which will be answered incorrectly or merely guessed at by those who have not been instructed. The same items should be very easy for those who have been instructed. Now, if the change from merely guessing before instruction to thorough knowledge after instruction occurs more clearly in one group than another on some items of a test, it could indicate that these items are biased against the group which shows less change. It may, of course, be that the instruction was at fault; in that case the curricular design should provide for that. Green (1975, page 50) recognizes that this method is "clumsy and messy," and, as far as can be ascertained, it has never been used.

Jensen (1976) has tried an experimental method of having judges from minority and nonminority groups (black and white psychologists, in his case) try to sort the eight most racially-discriminating items and the eight least racially-discriminating items into separate piles when all sixteen items were presented on separate cards randomly shuffled. It was found that the judges sorted no better than would be expected from sorting by chance. The items were from the Wonderlic Personnel Test and were very heterogeneous. Sandoval and Miille (1980) used college students (black, Mexican American, and Anglo American) as judges. They were not able to detect WISC-R items that are more difficult for a minority child than for an Anglo American child. The ethnic background of the judge made no difference in accuracy of selecting items that would be difficult for minority children. It might be that this method would be more effective with school-achievement items. It certainly should be more effective than this study indicates if they are correct who claim after casual inspection that many items in common tests are biased and that this is the cause of group differences in mean scores.

Those who claim that group differences are due to biased items, and that they are competent to merely look at items and determine which ones are biased (Williams, 1970, 1971), should be able to identify biased items.

These items should be the same items detected by methods which search for biased (or quantitatively discrepant) items. There should be a large number of these biased items, because a large number of them would be required to make the substantial differences in group mean scores that are found. Presumably, since during the development of most modern tests glaringly biased items are eliminated (if they are even written), it requires some insight and skill to detect among the remaining items those which are biased. It may be the case that people with those skills can write additional similarly biased items that will escape elimination during the pretest screening for bias. Possibly writers of deliberately biased items would produce different biased items for different minority groups. It would seem likely that those who could produce these kinds of biased items could also produce or identify unbiased items which were measures of the achievement of interest. If all of these assumptions are sound, then persons with the appropriate skills might be able to develop sets of biased and unbiased items which could be used in biased-item search studies much as marker variables are used in factor analysis. The procedure would be to include in a new test several biased items for each important minority and several unbiased items for that minority. Then a discrepant-item search technique would be applied to all the items. The new items which are biased should yield results similar to the deliberately-biased items, and the new unbiased items should yield results similar to the deliberately-unbiased items. If bias is a dichotomomous phenomenon, the results should be clearcut; if bias is a matter of degree, the results should indicate what degree of bias in an item should be of concern.

Another suggestion could be included in such a study. Angoff (Note

16) proposed that the standard for regarding items as biased could be set by using two extremely different groups of people from one culture in a discrepant-item search technique. The degree of discrepancy represented by the most discrepant items for these two similar groups would clearly not represent bias. Any items which were even more discrepant when comparing two different cultures would be suspect. The results from using this approach to setting the level of discrepancy that suggests bias could be compared with the levels reflected in use of deliberately biased and unbiased items.

While studies such as those described in the paragraphs above would be of great interest, they might be very difficult to conduct. First, it might be that no experts can be found who can recognize or create items in that gray area of being so strongly biased that they cause the large differences in group mean scores but so subtly biased that reviewers during pretesting and item editing will not detect their flaws. Jensen's psychologists and Sandoval and Miille's college students were not able to sort items into biased and unbiased categories, but perhaps other specialists could. Second, it might be difficult to perform some of the above experiments in a realistic setting, such as a state-wide testing program, a college-admissions testing program, or some other important procedure using test results. If forewarned of the experiment, examinees' behavior will change, but if not forewarned, it might be argued that they had been treated unethically.

An experimental method has been used which applies primarily to reading-comprehension items. In this method (Green, 1975; Tuinman, 1973-4;

Pyrczak, 1974), one compares the passage dependency of reading-comprehension items for different cultural groups. Pyrczak, Tuinman, and others have found that with reading-comprehension tests in which examinees are supposed to answer questions about a paragraph after reading it, many of the questions can be answered by a large proportion of examinees without their having access to the paragraph at all. Tuinman (1973-4, p. 219) gives an index of passage dependency as:

$$PDI = 1 - d_{np}/d_p \ ,$$

where PDI is the passage-dependency index,

d_{np} is the proportion who get the item correct without seeing the passage, and

d_p is the proportion who get the item correct after reading the passage.

Apparently Green would have the bias investigator conduct an experiment on reading-comprehension items in which different cultural groups were involved and random samples of each attempted to answer questions about a paragraph, some having seen the paragraph and some not having seen it. If the PDI's differed for the various cultural groups, the item would be regarded as biased. This method has never been tried out, and the standards for how large the sample should be, the statistical test for differences in PDI that should be used, and similar important details are unexplored at present.

A fifth experimental method was suggested by Jensen (1976) as an important check on whether bias, when found, was distinctly cultural. Jensen found that black and white children gave different proportions of responses to the various wrong alternatives in Raven's Progressive Matrices test. This at first seemed to indicate that blacks somehow interpret these items differently from whites. However, he then analyzed the responses from white children two years younger than the white children in his initial sample and found that the distributions of wrong responses for blacks were like the distributions of wrong responses for white children two years younger. Miele (1979) was also able to eliminate group differences in WISC scores of Georgia school children by comparing majority children with minority children one grade ahead. This sort of finding seems to shed a different light on the interpretations of the nature of the bias in what seemed to be biased items. However, it is unclear if there is a bias against a culture which happens to be black or against people who happen to perform at a maturity level one or two years behind another group of people. It is also not clear that a sound procedure would be to eliminate from a test those items which reveal this kind of an educational or achievement difference. But it might be important to include such items and analyze for discrepant items to determine in which kinds of content maturity was lagging and for which students. The treatment, as with the doctor's treatment of underweight for undernourished children, is not to revise the scale so that the lack of maturity will not show, but to teach the children so that the lag in maturity does not occur. This experimental procedure, then, might sometimes assist in differentiating for discrepant items between bias in the items and differences in level of learning.

Construction methods to insure against biased scores

Ebel (Test bias fears, 1975) states that there is no direct evidence that achievement tests commonly used in this country are biased, and it is improbable that such evidence will appear. He says that there are three steps in building an unbiased achievement test: (1) define a domain of tasks, (2) sample from that domain, and (3) calculate the reliability of the scores from that test on a representative sample of the population for whom the test is designed. If steps 1 and 2 have been done and the reliability is satisfactory, Ebel feels that the test will be reasonably unbiased. In measurement terms, Ebel has said that a test which is content valid and reliable will be reasonably unbiased. Therefore, to make unbiased tests, make content valid and reliable tests by the usual test-construction techniques.

Angoff (Note 1) describes a more elaborate scheme of the same general nature to deal with the problem of achievement tests for school subjects in which different specialized curricula have been developed, such as the three versions of the Biological Sciences Curriculum Study (Blue, Green, and Yellow) developed in the 1960's. If a test is to be developed that will not favor one curriculum over another, how is it to be done? Angoff suggests that one isolate a common core of knowledge and abilities basic to all the curricula, and write items to measure this core. Then get expert judgments of the degree of bias for each of the common-core items. Follow these with the use of plots of item difficulties (or some other discrepant-item search technique), so that any items which are biased can be eliminated from the core. Then develop items which will measure the content that is specific to each of the various curricula. Combine the common-core unbiased items with the curriculum-specific items in a balanced manner into a single test, and give it in its entirety to students enrolled in the various curricula. Use these students' scores on the common items (and their scores on any available common reference-ability test, such as the SAT) to adjust and control for differences in the abilities of the students enrolled in the various curricula in an analysis of covariance. That analysis is to determine whether the test as a whole is unbiased or still yields higher or lower scores for any subgroup than their abilities would warrant.

There are several significant aspects of this method that should be pointed out. First, like Ebel's process, it demands content (and Angoff would add, construct) validity. Second, it does not look for a test composed only of unbiased items. Instead, it seeks a core of unbiased items and a balance of biased items. Third, it makes use of the unbiased score on core items to evaluate the fairness of the balance. This method does require that judges be able to evaluate the fairness of items, something which some of the other methods have called into question. Angoff emphasizes, however, that the judges are not entirely trusted; after their judgments have been made, a discrepant-item search technique is used to verify that they have not let biased items slip through. Unfortunately, the discrepant-item search techniques available have not yet been thoroughly proven, as we shall see. Angoff, of course, recognizes the imperfection of this procedure.

Miscellaneous methods

Several methods do not readily fit the classifications we have used above. Jensen (1980) suggested a psuedo-group method to determine whether discrepant items are due to cultural bias or to some other variable. He suggested forming a group from the majority group that conforms to the distribution of total test scores in the minority group. The discrepant-item detection technique is then applied to this psuedo-group and the total group from which it came. There are by definition no cultural differences between these groups. If the same items appear as discrepant between these psuedo groups as appear discrepant between the majority and minority groups, the bias is obviously not cultural but is due to ability level differences between the majority and minority groups. This same approach can be used to detect whether a bias is due to socioeconomic level (instead of ability level) or other suspect factors.

Ozenne, Van Gelder, and Cohen (Note 41) used a procedure (described by Merz, 1977) for debiasing norms for a test. The method consisted of plotting item difficulty values of cultural groups vs. a reference group such as the published norms groups. Items are first arranged in order of difficulty for the reference group. These item numbers are put on the ordinate of the graph, and the new group difficulty values are plotted on the abscissa. If the "uniformity of the shapes of the curves was disturbed" (Merz, 1977, p.14), visual inspection reveals the biased items. Then the point-biserial correlations between item score and total scores are computed for the new group. Items with low item-total score correlation for the group are examined for bias. Expert judgment is used as the criterion. According to Merz (1977), the method used to identify an item as biased was not clearly described, and the set of decision rules generated by the experts or used by them was not specified in the report of the use of this method by Ozenne, Van Gelder, and Cohen.

The last method we will describe was also presented by Jensen (1980). It compares differences in item difficulties between groups with the differences that would be expected if each group were compared with itself. Difficulty values are first transformed to normal deviates; then the normal deviates from one group are transformed to the mean and standard deviation of the other. An ANOVA of the items-by-subject matrix of each group is performed to obtain the average reliability of the item difficulties, and the standard error of the difference between the delta values is calculated. These elements are used in a chi-square test for each item. Items with significant chi-square values are suspected of bias. Jensen (1980, p. 441) says, "A first step toward improving a significantly biased test is to eliminate all items with chi-square values greater than 3.84."

Applications of the methods

In the previous section of this paper, we described a large number of different methods of locating and identifying discrepant items. Some of those methods have been tried out on data. Single methods have been explored, sometimes on artificially generated data, sometimes on real data. Several different methods have been used on the same data to determine whether they agreed in locating discrepant items. This approach has also been used both on artificial data and on real data. Some of the methods have actually been applied during test development; items have been eliminated from tests because they were discrepant. In fact, whole categories of items have been eliminated from tests, on occasion. In the next section of this paper, we will examine these applications of the techniques for locating discrepant items which might be biased. We will find that: (1) little or no evidence indicates that flagged items are actually biased due to a clear irrelevancy, (2) little or no evidence shows that mean differences between groups disappear when flagged items are removed, and (3) agreement among methods is limited, and little or no evidence indicates that one method is superior to another at flagging items which are biased due to irrelevancy and which account for the differences between means of cultural groups.

Empirical tryouts of single methods

TID methods

Hills and King (Note 42). Hills and King used a form of the TID method as an approach to detecting items which should be examined by judges for possible bias. They did not convert p values to the College Board's delta scale, arguing that normalizing the scale would effectively serve only to emphasize differences in p values that were near 1.00 since the tests under consideration, the Florida Statewide Student Assessment Tests, are minimal competency tests and therefore have few if any difficult items. They did attempt roughly to equalize the groups for ability by obtaining data for ten times as many whites as blacks, dividing the blacks' score distribution into ten or eleven intervals, and choosing at random from the same intervals among white students a number equal to the number of blacks in each interval. That same procedure was followed for Hispanics, and it was used in equalizing the score distributions for males and females. They also checked the test data to be sure it satisfied Reckase's criterion for unidimensionality.

Hills and King did not calculate the major axis of the ellipse. Instead they examined the plots and drew a line through approximately the center of the ellipse, a line which usually was close to the 45 degree line since the groups were equalized roughly on ability. Potentially-biased items were those which were most distant from the hand-drawn line. The number of potentially-biased items to be examined for bias varied from plot to plot depending on the spread of the items from the center line and the appearance of points distinctly different from other plotted points (Sinnott, Note 5).

The Florida Statewide Student Assessment Tests provide scores for communications and mathematics. Hills and King studied the data from the 1980 administration of the two tests to eleventh grade students. One of the tests at that level concerns basic skills in communications and mathematics. The other concerns applications of those skills to practical problems in everyday life. The smallest number of cases used for any group was 1,000; the largest, 7,500.

After potentially-biased items had been selected based on the plots of item ps, eleven judges representing whites, blacks, Hispanics, males and females were asked to explain the reasons for the item discrepancies. The judges were given sets of items to evaluate, each set containing some items with discrepant p values and other items with concordant p values. They were first to indicate which of the items were discrepant, then which group found the items relatively easier, and finally why. The judges were all experienced public-school teachers or administrators.

A total of 82 items were selected from the plots as being discrepant. The judges agreed among themselves on only 2 communications items which they said favored whites compared to blacks, 7 communications items and one mathematics item which they said favored whites compared to Hispanics, and on one communications item which they said favored males compared to females. However, they often identified concordant items as discrepant or misidentified the direction of discrepancy. Often they identified one item as biased but did not identify as biased another item adjacent to the first in a scatterplot. The reasons the judges gave for items being biased were so inconsistent that there is no assurance that the items that they identified as biased and that were discrepant in the plots are really biased at all.

Hills and King concluded that the method they used may work very well for detecting potentially-biased items, but the test to which they applied the method had with so little bias, if any, that it was not an adequate test of the method. The real problem is to find items that are sufficiently biased to produce large discrepancies in p values which are readily explainable by judges afterwards but whose bias is not detected by the reviewers who screen items for bias before the items are assembled into operational test forms. Such items, if they can be found, could be inserted into tests as "marker items." They would provide the boundaries in the scatterplot. Other items which fall outside those boundaries would be potentially-biased and have some chance of being biased items that would be recognizable, explainable, due to irrelevant factors, and removable from a test with some chance of reducing differences between mean scores of different cultural groups.

Rasch residuals

Wright, Mead, and Draba (Note 18). Wright, Mead, and Draba (Note 18) described in one paragraph their analysis of the responses of white, black, and Spanish-speaking students using their method of comparing differences in residuals after fitting the data to the Rasch model. There is no description of the test that was used except that it was a well-known reading-comprehension test. They found that the residuals tended to be positive for items based on poetry for blacks compared to whites. Residuals were negative for natural science passages.

ICC goodness of fit

An application of Durovic's method is described very briefly by Scheuneman (Note 6). Durovic reports applying his test of whether the fit to the one-parameter ICC model differed for two groups on a 14 item test. He found two items discrepant. Those two items were also objected to by two minority group reviewers who had no knowledge of the statistical analysis results. Since the kinds of test development procedures now widely used would be expected to eliminate flagrantly biased items before test publication, one wonders whether the Durovic procedure would detect items with less extreme problems.

Three-parameter ICC

Lord (1977). Lord (1977) studied bias in the items of the 85 item verbal section of the SAT using the three-parameter ICC model. The subjects were 2250 whites and 2250 blacks, both groups being about 44% male. The procedure tested whether the a and b parameters (discrimination and difficulty) were the same for blacks and whites. It turned out that the null hypothesis was rejected at the .05 level for 38 of the 85 items, i. e., 45%. Several items had crossing ICC's, a kind of discrepancy that would not be detected by a method such as the TID procedure. Examination of the items failed to yield any insight into the reasons for the discrepancies of individual items.

Linn, Levine, Hastings, and Wardrop (1980). Linn and his colleagues tried the ICC method on data from the Anchor Test Study (Bianchini & Loret, 1974). They obtained item response data on the Reading Comprehension section of Form F of the Metropolitan Achievement Tests for students in grades 5 and 6. Using about 2000 cases in each of 8 groups based on income of parents (2 levels), grade (5 or 6), and race (black or white), LOGIST yielded ICC parameters for each group. Linn and his colleagues developed approximate standard errors of the ICC's and plotted the ICC's with their standard errors for pairs of groups. They also calculated areas between ICC's, the area between two ICC's in which one comparison group was above the other over a large part of the ability range, and the square root of the sum of the squared differences between ICC's over the middle ability range.

Item Bias

When white students were compared with each other across income or grade level, there were smaller differences than when whites were compared with blacks. However, when blacks were compared with each other, the differences were also larger than when whites were compared with each other. The agreement among the four independent between-race comparisons on which items are discrepant is far from perfect. However, three of the 45 items were identified in all four pairwise comparisons between races. On all three of these, the ICC's cross. One of the three seems to favor predominantly blacks because the ICC for blacks is above the ICC for whites in the low ability range where most of the blacks fall. Another seems to favor whites for the analogous reason. Examination of the content of the items did not indicate what would produce the discrepancies, and items of similar content were not always discrepant. The elimination of the three discrepant items would have trivial effect on mean differences in scores between blacks and whites.

Linn and his colleagues concluded that examining the content of the items to detect bias is unsatisfactory, and the different white-black comparisons used in this study did not agree particularly well. They provided an illustration suggesting that simply comparing differences in ability parameters in the three-parameter model, or comparing discrimination parameters, is unsatisfactory because large differences in them can occur in situations in which the differences in ICC's are very small over the entire ability range (theta ranging from -3 to +3). Linn and his colleagues argue that the entire ICC's must be examined to locate soundly the discrepant items.

Linn and Harnisch (1981). Linn and Harnisch (1981) tried out their method of comparing estimated probabilities of correct responses based on the three-parameter logistic model with the observed frequencies of correct responses for various subgroups. The test they used was the mathematics part of the Illinois Inventory of Educational Progress (IIEP) given to a sample of eighth grade students. Their illustration focused on a group of black students, but they indicate that similar analyses (unreported) were made for other subgroups. All but one of the 46 test items were studied. One item had a discrimination parameter near zero and was left out. LOGIST was used with all 2,055 students with usable data. The number of black students was 283.

The blacks were divided into fifths based on estimated abilities. Within each fifth, the observed frequency of success and the probability of answering the item correctly based on their ability estimates were compared. These values were graphed. Average standardized differences between the observed values and the probabilities were calculated as an overall index of discrepancy for an item. Linn and Harnisch present the graphs for the four most extreme items, two deviant in favor of blacks and two in favor of whites. They sought for clues as to causes for the differences in terms of content and format. They found that blacks had greater difficulty on items dealing with the metric system, but they had less difficulty than whites on story items dealing with money. Since both kinds of items are appropriate content for a mathematics test, Linn and Harnisch viewed this as not very helpful, though they note that adjusting the proportions of these kinds of items in the test would alter the rate of passing the test or the group differences on the test. However, they question whether the items are at fault and suggest that it is just as likely that other factors caused the discrepancy, i. e., the ICC model might be inappropriate and unidimensionality of these questions was doubtful. Further, Linn and Harnisch speculate that there may have been differences in the amount of instruction given in these kinds of problems to black and to white students by virtue of school to school differences in attendance patterns and content coverage. It seems just as plausible, however, to suspect that the differences are not in instructional emphasis but in the motivation to learn from the instruction that is provided. Some cultural groups may be more highly motivated to learn abstract concepts that appear to have no immediate application in daily life. This hypothesis in one form or another has often been invoked to explain the high level of success of Orientals and Jews on academic tasks. (See, for example, Sowell, Note 43).

Rudner (1977a). Rudner (Note 3) studied the three-parameter ICC approach in which the indicator of discrepancy was the size of the area between ICC's for different groups. His subjects were normal-hearing and hearing-impaired students. The test was the 1973 Stanford Achievement Test, Form A, Primary 2 Battery, Reading Comprehension Subtest. It consists of 16 paragraphs with a total of 48 four-choice items.

Nearly 2500 hearing-impaired examinees were divided at random into two groups. The first comparison in this study was between the ICC's developed separately from these two psuedo-cultural groups. It would be expected that most of the differences between ICC's in this comparison would be small since there can be no biased items in this situation where there is no cultural difference. Then one of these groups was compared with a normal-hearing group to examine the sizes of the differences in ICC's between groups that might be expected to be culturally different and for which some items might be biased. When groups from the same culture were compared, it was found that several items did not fit the model. Two items had large enough differences in ICC's (area of difference between ICC's of 0.4 or greater) that they appeared to be discrepant even when there were no differences between groups. Rudner attributes these results to local dependence, multidimensionality within groups, or poor parameter estimates. When different groups were compared, 10 items (22%) had areas between ICC's of 0.4 or more, a size Rudner regarded as significant. In at least one case, ICC lines crossed; the item favored one group at high levels of ability but the other group at low levels of ability.

Confirmatory factor analysis

Rock and Werts (Note 19) applied the confirmatory factor analysis technique to the SAT. The verbal section of that test contains four item types, the mathematical section, two. If factor patterns based on these kinds of items can be fitted from the data, then one can study the invariance of the patterns across cultural groups. The cultural groups used in this study were American Indian, black, Mexican American, Oriental, Puerto Rican, and white. Five hundred cases from each group were studied, using the computer program COFAMM to make the various statistical tests.

Although with these large samples the statistical tests were sometimes significant, the authors concluded that in all cases the sizes of the differences among groups were trivial. The factor patterns, the scaling units, and the standard errors of measurement were judged to be equal for all size groups. Thus, no psychometric bias was evident. The same conclusions applied to both SAT-V and SAT-M. Reliabilities may differ across the groups, but primarily due to differences in variability in the groups. Similar analyses were done on the Verbal score, the Mathematical score, and the Test of Standard Written English score, without considering the item types within those tests. The results were the same as for the analyses within the Verbal and Mathematical tests.

Rock and Werts argue that it is customary in item-group interaction studies to define items that are exceptionally hard for some group as biased. However, information on covariances between items is not taken into consideration in those analyses. The finding of differentially difficult items may or may not indicate bias. It may imply differential rate of development rather than bias. When the covariance structures are invariant across groups, one can look at differences in factor means as evidence of different states of development. That kind of analysis revealed little in the way of interaction between cultural groups and kind of item within the Verbal or the Mathematical tests. But a striking interaction was observed in factor means across the total scores on SAT-V, SAT-M, and Test of Standard Written English. The total factor mean scores of Orientals on SAT-M are nearly three-fourths of a standard deviation higher than their total factor means on the two other scores. A similar but less extreme interaction is observed for Puerto Ricans. Within other cultural groups, the total factor mean scores on the three variables are much more nearly equal. The interactions are interpreted as differences in development rather than item bias.

Norms debiasing

Bianchini (1977) reported an application of the Ozenne, Van Gelder, and Cohen method to the problem of debiasing norms for students in "minority-isolated" schools. The test was the Reading Test of the *California Achievement Tests,* 1970 edition. The method Ozenne and his colleagues used detected 13 of 82 items as discrepant. These items were removed from the test-scoring procedure, and the remaining 69 items were rescored. The results were quite disappointing. On the biased test, the average percent correct was 49 on the national sample; it went up to 51 on the debiased tests. For the minority sample, on the biased test the percent correct was 35; on the debiased test it rose to 37. So the minority group was no better off on the debiased test than on the biased test. The GD/I ratio proposed by Jensen must have been high, something which could have been examined before proceeding with "debiasing" the norms.

Empirical comparison of methods using computer-simulated data

ICC vs. TID and other methods

Groome & Groome (Note 7). Groome and Groome (Note 7) described a study in which they compared a three-parameter ICC approach with a TID approach when the difficulty parameters of the items were different for two groups, and the item parameters were either normally distributed or uniformly distributed. Their description of data generation is not clear, nor is it clear which item parameters were distributed normally or uniformly or why they regard the distributions of item parameters as important. Their simulation used generated data for 1000 cases in each of two groups, with the subjects distributed normally and with equal ability in the two groups. Each test was designed to be 60 items long. Four simulated tests were used, one with item parameters distributed normally and no discrepant items, another with item parameters distributed normally and some discrepant items, a third with item parameters distributed uniformly and no discrepant items, and the fourth with item parameters distributed uniformly and some discrepant items. The method for injecting discrepancy into some items is not described, but it appears that after the data were generated the Groomes used the ICC three-parameter model to decide which items were discrepant. Then they evaluated whether the TID method detected the same items. Rudner's approach of evaluating the area between ICC's was used, with an area between the curves of .70 being the criterion for calling items discrepant.

On the concordant items, using a criterion of 1.5 times the standard deviation of the residuals, the TID method called 6 items discrepant when the item parameters were distributed normally and 8 when distributed uniformly. With a .95 confidence interval as the criterion for discrepancy, no items were detected as discrepant among these items designed to be concordant. On the items designed to be discrepant, the same number were detected as discrepant using the 1.5 s.d. criterion. Using the .95 confidence interval, only one item was detected as discrepant. That single discrepant item was identified when the item parameters were distributed normally. Since there is no difference between the discrepant and the concordant pairs of tests in the number of items called discrepant, one is disappointed in the TID method. One is dismayed to find that with the tests containing items designed to be discrepant, the ICC method detected 7 items as discrepant when parameters were distributed normally, but the TID method only detected four of these while incorrectly identifying 2 others as being discrepant. When item parameters were distributed uniformly, the TID method only detected 6 of the 14 items the ICC method determined to be discrepant. At the same time, it again falsely identified two items as discrepant. These results occur in spite of the fact that the size of the residuals in the TID approach and the size of the area between ICC's correlate .95 when item parameter distributions are normal and .90 when those distributions are uniform.

Rudner, Getson, & Knight (1980a). Rudner, Getson, and Knight (1980a) studied three methods, with variations on each of them, using computer-generated data. They used TID, ICC, and C2 approaches, with two variations on the TID method, three on the ICC method, and two on the C2 method. The variations on the TID method were that one approach took deviations from the 45 degree line (using deltas standardized within groups) while the other took deviations from the major axis (unstandardized deltas). The variations on the ICC method were that one approach used the areas between the curves as the measure of discrepancy, another used the differences in degree of fit to the Rasch model, and the third used differences in the easiness parameters. (Both discrimination and difficulty parameters were designed to differ among items, so the fit to the Rasch model would not be expected to be good in some contrasts.) The variations on the C2 method, which was patterned after Scheuneman rather than Green and Draper, were that one approach permitted only 5 intervals while the other allowed the number of intervals to vary depending on the characteristics of the data.

The simulated data were designed to have different discrimination and difficulty parameters for different groups, and the amount of variability in these parameters was also varied from zero to a substantial amount. Thus some comparisons involved tests whose items differed only in discrimination, while other comparisons involved tests whose items differed only in difficulty, with many other combinations also being examined. Seven different test lengths were studied with four degrees of discrepancy in discrimination and four degrees of discrepancy in difficulty. Each set of data provided an N of 1200 cases in a group, and the groups were designed to have normal distributions of ability with means one standard deviation apart but equal standard deviations. Obviously, this was a complicated study.

The results do not provide much encouragement that these methods will agree in detecting discrepant items. For example, the correlations between the amount of generated discrepancy and the amount of detected discrepancy ranged from 0.46 to 0.69 if we exclude the agreement between the ICC method used to generate the data and the use of that same method to detect the discrepancy that was generated. The TID method that judged discrepancy by deviation from a 45 degree line gave the best results on the average except for the method which was used to generate the data. (That this method is not precisely equivalent to the usual TID method which uses distance from the major axis of the plot of transformed item difficulties as the indicator of discrepancy is evident in the correlation between the two methods, a value of .78.) The ICC method which examined fit to the Rasch model gave the worst average results. When there were no differences in difficulty of the items (all differences were in discrimination), the C2 method with 5 intervals was most effective, much more effective than the next contender which was the method used to generate the data. When there were no differences in discrimination values (all differences were in difficulty), the best method was the TID method using deviations from the 45 degree line. One might have expected a method based on the Rasch model to shine here, but the method based on comparisons of goodness of fit was the least effective.

Correlations between the methods were disappointing, too. They ranged from a low of .29 to a high of .87 between the two C2 techniques.

These results are confusing. A method which should work well in this situation does not. The method used to generate the data might be expected to correlate very highly or even perfectly with the results but does not, and the methods correlate with each other as low as a value of 0.29. Such results might cause one to hesitate to trust any of these methods as a practical way to locate and remove biased items from a real test.

Merz & Grossen (Notes 39 and 40). Merz and Grossen (Note 39) reported to an annual meeting of the National Council on Measurement in Education a study of six methods of examining test item bias. The methods were the TID, the item selection method, the chi-square method, a principal components factor analysis, and two variations of the ICC approach, the one-parameter and the three-parameter models. Two groups of simulated subjects were created, with about .4 s.d. difference in ability. The results reported at the meeting were quite confusing; a later final written report (Note 40) on the grant supporting this research eliminated the aspect of groups' differing in ability level, but reported on the same methods used in the earlier study. Data sets were created for two groups of equal ability; half of the discrepant items favored each group. Data were generated to simulate two levels of test difficulty, an average of 60% correct and an average of 80% correct. Within each level of difficulty, three levels of difficulty discrepancy were introduced, no discrepancy, 10% of the items discrepant, and 20% of the items discrepant. The amount of difficulty discrepancy in discrepant items is not discussed; a discrepant item is one in which the area between ICC's is greater than .70. The data were generated based on the three-parameter ICC model varying only the difficulty parameter. One thousand cases were generated for each of the 6 groups, and the tests were simulated as being 60 items long.

In Merz and Grossen's TID approach, item p values were computed within each of two groups being compared. They were converted to z scores within the groups (which were of equal ability), and then to the delta scale. Since the groups were of equal ability by design, Merz and Grossen used distance from a 45 degree line through the plot of points as the discrepancy indicator. These values were again converted to z scores.

In the item selection method, point biserial correlations were computed between item scores and total scores within each group of subjects. The difference between these correlations was the measure of discrepancy. These differences were converted to z scores.

The chi-square method used here was the standard chi-square but with a sign giving the direction of the difference between groups. At each ability level the observed-minus-expected-difference was evaluated. If the number of observed correct responses was smaller than expected for a group, the item was considered to be biased against that group at that ability level and was given an appropriate sign. These signed values were added across ability levels and groups to give the signed chi square for the item. Obviously, if an item favored one group at some ability levels and the other group at other ability levels, the resulting chi square would be small, and the discrepant effects would be hidden. Six to eight intervals were used to group the scores. The signed chi-squares were used as the measure of discrepancy, and they were converted to z scores.

The factor analysis method employed by Merz and Grossen used group membership as a dummy variable. Principal components analysis was used with varimax rotation. The bias factor was identified as the one with the largest loading for the group membership variable. The bias of an item was its factor loading on the biased factor. These bias values were converted to z scores.

The three-parameter ICC approach used Urry's (Note 44) method for estimating parameters. The areas between ICC's were computed, and converted to z scores. For the one-parameter model, the parameters were estimated using the BICAL program (Wright, Mead, and Draba, Note 18). Areas between ICC's were calculated and converted to z scores.

Merz and Grossen state (Note 40, page 41), "In order to determine how well each of the six methods detected bias, the standard score conversion of the measure of bias for each was correlated with the actual degree of bias generated in the data." It is not clear to the reviewer what variables were correlated. We have z scores for individual items reflecting their degree of discrepancy or bias, and we have amounts of discrepancy introduced into tests in terms of percent of items which are discrepant not the degree of discrepancy of individual items. It is not clear how the indices for bias of individual items were converted into standard score conversions of measures of bias for methods. Therefore, from the report we cannot determine what variables were correlated. As a result, it is not clear what the results and conclusions of this study signify. However, their reported results indicate that for most sets of data, the ICC methods correlate substantially with the data which were generated using ICC methods. The failures are for easy tests with many discrepant items (average item p value of 0.80 and 20 percent discrepant items). By comparison, the item selection method had relatively low correlations with the generated amount of discrepancy. The correlations with the TID method were very high for all sets of data (at or above .95), and they were substantial for the chi-square method (.75 to .91). The factor analysis method, like the ICC methods, worked well for all but one group, but in this case it was the difficult test with only 10 percent of the items discrepant that presented the problem.

When the methods are evaluated in terms of whether they accurately identified the specific discrepant items, and whether they erroneously identified other items as discrepant also, the results are not as promising. Most of the methods tend to identify more items as discrepant than were generated to be discrepant and often they identify the wrong items. The worst case was the factor analysis method which identified 7 items as discrepant on the difficult test with only 6 discrepant items. However, none of the 7 identified as discrepant were items created with difficulty discrepancy in them. In general, it appears that about half of the items identified as discrepant were those created to be discrepant. One certainly would not be comfortable in excluding items from a test based on the results of this study. One might eliminate items identified as biased but find that he retained many of the discrepant items and discarded perfectly sound items.

Empirical comparisons of methods using live data

Seven studies have compared various methods of detecting bias using test data from live examinees, but with no apparent intent to use the results for anything but a comparison of the methods. Two of them included very similar sets of methods, the TID, the ICC, the C2, and an item selection method. The third used a factor analysis method instead of the item selection method, but the other methods used were TID, ICC, C2 and chi-square. The fourth and fifth studies used these methods. Another study compared two ICC methods, and one compared ICC person-fit methods with distracter-analysis methods. The studies used different test content and different kinds of examinees.

TID, ICC, and C2 with other methods

Rudner & Convey (Note 17). Rudner and Convey (Note 17) reported a study using the 1979 edition of the *Stanford Achievement Test*, Primary 2, Reading Comprehension with hearing-impaired versus normal-hearing students. The test contains sixteen paragraphs with 48 four-choice items. (Items of this kind may violate the assumption of linear independence required by the ICC technique.) There were 2,637 hearing-impaired subjects and 1607 normal-hearing subjects. The hearing-impaired were divided into two groups so that a psuedo-group comparison could be made. The same data were used in another study of the ICC method by itself, discussed earlier.

One of the methods used in this study was the TID method with deviations taken from the 45 degree line of plots of delta values standardized within groups. This method should not be expected to work well if there are real differences in reading comprehension of the groups. The second method was Scheuneman's C2 technique with five intervals. In addition to using the size of C2 as an index of degree of discrepancy, Rudner and Convey used one minus the probability associated with a chi-square that would correspond to the C2 index. (This study was done before publication of the papers showing that the C2 technique was not related to chi square in a way that would allow probabilities from chi-square tables to be used with the C2 technique.) The three ICC item parameters for each item were estimated separately for the two groups involved in a comparison using a chi square procedure proposed by Urry. The parameter value estimates were equated, and the areas between estimated equated ICC's were used to indicate the degree of discrepancy.

The fourth method was based on factor analysis. This method proceeded as follows:

a. The item responses for the two groups being compared were combined.

b. Inter-item product-moment correlation coefficients were calculated.

c. The matrix of correlations was factor analyzed by the principal components method extracting factors until an eigenvalue of one was reached.

d. The factor matrix was rotated to orthogonal simple structure by the varimax criterion.

e. Factor scores were computed for each examinee on each factor.

The method of computing factor scores is not described in the report. To determine which factors are biased, each group's mean factor score for each factor is calculated and the differences between means for the groups being compared is tested with a t test. If the difference is significant at the .001 level, the factor is called biased. The degree of bias of an item was indexed as its largest loading on a biased factor. This method seems to correspond in part to the method suggested by Merz which was to incorporate analyses of variance of factor scores for examinees.

(Please continue on the next page.)

The Rudner and Convey study had an interesting facet in that the hearing-impaired group was divided into two subgroups of approximately equal ability. The analyses were run on these subgroups which have the same culture. This is similar to Jensen's psuedo-group method, which is used to be certain that a "bias" is really a cultural bias. In Jensen's method, however, persons are taken from a cultural group in such a way that the subgroup has the same distribution of scores as the group of the different culture. (To use Jensen's method, Rudner and Convey would have chosen a group from the hearing-impaired that had a mean of 28.9 and a standard deviation of 12.44 to match their normal group's characteristics.) The results from the equal ability subgroups in this study can be considered as a base line against which to compare the results from groups whose cultures differ instead of as a test of whether observed bias is cultural.

The factor analysis results from the two situations are interesting. Only 7 factors were found in the analysis for the groups which differed. Six of those factors were found to be biased. For the groups which did not differ--which were random samples from the hearing-impaired population--14 factors were found, but only 3 were found to be biased. Some of the loadings on these factors, which are the measure of amount of bias, are as high as any of the loadings for the analysis of the groups whose cultures were thought to differ. By definition there cannot be culturally biased items if the groups are from the same culture, but with two groups from the same culture the method found biased factors and found items with large loadings on those biased factors. By contrast, the C2 and TID methods found no discrepant items in the equal-groups comparisons. These results inspire no confidence in the factor analysis method used here.

The C2 method ran into the problem of extreme differences in distributions of ability between the two different groups. Highly disproportionate numbers of hearing-impaired occurred in each interval.

Agreement among the methods was only modest on the data from the diverse groups. The correlations between bias scores ranged from .17 to was smaller for the different groups than for the groups designed to be the same. There are marked differences among the methods in number of items identified as biased. All but one are identified as discrepant by the C2 method using 1 - p as the measure of discrepancy. The ICC method found only 5 items in the diverse groups with discrepancy indices greater than the largest index for the similar groups. The authors suggest that the ICC method may not have been appropriate for these data since there were only 45 items, while a minimum of 60 may be needed.

Not only did the methods not agree with each other, and disagree on how many items were biased, they failed for the most part to agree on which items were biased. Using arbitrary decision rules (which appear to the writer not to reflect sound use of the equal-groups results as a comparison basis), the authors cross-tabulated the items and the methods in terms of detecting bias. Only three items were agreed on as being biased by three of the four methods. Seven more were identified by at least two methods as being biased. Seven more were identified by only one method.

Results such as these are very discouraging. But one must realize that the TID approach may not have adjusted properly for ability differences, the ICC method was used when it was suspected that there were not enough items and when it was known that the items were not linearly independent, the C2 approach had problems setting up useful intervals since the groups differed in ability, the 1 - p approach was based on an erroneous assumption about probabilities associated with C2 values, and the factor analysis approach used factor methods that were inappropriate. The theoretical analyses presented at the beginning of this paper might have suggested that results such as these would be expected from these methods.

Ironson & Subkoviak (1979). Ironson and Subkoviak (1979) used three of the same methods used by Rudner and Convey, but used item selection methods in place of the factor analysis approach. Their subjects were 1691 blacks and 1974 whites tested as part of the National Longitudinal Study (NLS) (Hilton & Rhett, 1973). The tests were from the NLS battery; six subtests were studied. Three are called traditional tests: vocabulary, consisting of 15 items on synonyms, reading, consisting of questions based on passages of interest to minority students, and mathematics, consisting of quantitative comparison items. The other three are called nontraditional: picture-number, an associative memory task involving recalling numbers associated with pictures, letter groups, an inductive reasoning test based on groups of letters and determining which group is different from the rest, and mosaic comparisons, a test of perceptual speed and accuracy. (Since the number of items in each test is not reported, one conjectures that problems might arise with estimating ICC parameters due to inadequate numbers of items in subtests.)

The TID approach used by Ironson and Subkoviak evaluated item discrepancy on the basis of deviations from the major axis of the points in the scattergram of transformed item difficulties. This is a more appropriate procedure for groups of differing ability than the procedure used by Rudner and Convey. The discrepancy index was given a positive sign if the blacks tended to get the item correct less often and a negative sign if whites were less often correct, as Ironson and Subkoviak used the method. Scheuneman's C2 was used, and a modification of the C2 technique was also used--a signed C2 similar to the signed chi square used by Merz and Grossen which was described earlier. The ICC approach used the area between ICC's as the measure of discrepancy. A signed index was used here also, with the area being added to the discrepancy index when the curve for the whites was above the curve for blacks, and subtracted from the index when the curve for the blacks was higher. The item selection method used here involved calculating the point-biserial coefficient of correlation between passing the item and total test score. This was done separately for blacks and whites. The unsigned index of discrepancy was the absolute difference in size of the coefficient for blacks and whites. The direction of discrepancy was preserved by attaching an appropriate sign to the difference.

Ironson and Subkoviak introduced a novel feature in their study. They arranged for a psuedo-group comparison as suggested by Jensen. The white group was divided randomly into two halves. One half was used to study differences between blacks and whites. The other half of the white group was used to form a third group with the same ability distribution as the blacks. For each subtest, a random sample of 100 blacks was chosen. From the unused half of the white group, a group of students was chosen who were matched on the basis of the subtest total score to the black sample. Using these matched groups, white minus black item *p* values were computed. These values were correlated with the signed bias indices from the rest of the data. (The correlations ranged from -.03 for the index based on discrimination differences to .41 for the C2 index.)

The results of the analysis are again not very satisfying. Some aspects encouraged the authors. They regarded as reassuring the fact that only 2% of the items were found to be discrepant using Scheuneman's C2 technique when two groups of white subjects were compared, but 24% were found to be discrepant when C2 was used to compare whites to blacks. However, we saw in the introduction to this paper that significance tests based on C2 are not to be trusted. The size of the discrepancy indices tended to be two to three times as great for black vs. white comparisons as for white vs. white comparisons, also, but, again, these are unmatched groups, and further there are regional differences as well as black vs. white differences. The blacks came from various parts of the U. S., but the whites came from the north central region only. It is possible, then, that the bias found was not due to ethnic group or to characteristics of the items, but rather to differences in curriculum, instruction, student motivation, or other factors.

Correlations among the methods are disappointingly low and erratic. For one analysis, the authors grouped all subtests together. (This should do violence to any method that assumes unidimensionality.) The correlations among methods in that analysis ranged from -.09 to .48 for unsigned bias indices and from .08 to .62 for signed indices. The mean correlation for the unsigned indices was .22. The mean correlation for the signed indices was .34. The lowest correlations involved the item selection method.

A wide variety of correlations occurred between methods across subtests. For example, for unsigned indices, TID correlated with the item-selection method indices from -.42 to +.21. The C2 and ICC unsigned methods correlated from -.05 to +.74 depending on which subtest was involved. For signed indices, similar variability depending on subtest was evident. The authors examined the 24 items they regarded as most biased to see whether the methods agreed on picking these items out. Agreement was modest among ICC, TID, and C2 methods, from 33% to 54%, but the item selection method agreed with no other method above the chance level. The ten most biased items as determined by each method were examined to see what in their content was causing the problem. The authors found it difficult to locate the source, but many of the items were verbal items and items at the end of the subtests. In studying bias, one should keep in mind that bias may not be in the items but in the test-taking behavior of the different cultural groups. If some groups tend to be slower than others, items near the end of a test may appear to be biased regardless of the content or form of the item.

Once again the results of this study do not point clearly to any simple conclusion. There are also some deficiencies in the report. For example: some of the analyses violated assumptions, there may have been too few items for the ICC method, no mention is made of the concern that should have occurred with the C2 method in having enough subjects at each level with groups of differing ability or how that problem was resolved, and the item-selection method seemed to be out of step with everything else. One is not assured that different methods will select the same items as biased, or that consistent results can be expected across different subject matter.

Shepard, Camilli, & Averill (1981). The third empirical comparison using real data was done by Shepard, Camilli, and Averill (1981). This study used the same methods that were used by Ironson and Subkoviak, but with some more variations. The subjects were groups of about 500 black, Chicano, and white students in grades, 4, 5, and 6. (The data were initially collected by Jensen for a different purpose.) The test studied was the Lorge-Thorndike Intelligence Test which has verbal and nonverbal scores. Raven's Coloured Progressive Matrices was also used in a special way in this study. Comparisons to reveal bias were made between whites and blacks and between whites and Chicanos, but not between blacks and Chicanos. The groups were not equated on ability.

The TID method used in this study followed Angoff's procedure. Perpendicular deviation from the principal axis of the plot of points was used as the index of discrepancy. A positive sign was given the discrepancy if the item was easier for whites; vice versa for blacks. The C2 method was based on Scheuneman's method for dividing the groups into intervals instead of Green and Draper's. In this study, both signed and unsigned C2 values were used. The signed indices are formed by inspecting the direction of discrepancy in each score interval and giving it a sign before summing discrepancies across intervals. Thus if ICC's cross, a lower C2 or chi-square will be found using signed values. A standard chi square was used, in both signed and unsigned forms, making a total of four different "chi-square" approaches.

Ten different approaches based on ICC's were used in this study. The LOGIST program was used to estimate three item parameters for each item in each group. (Note that the groups are rather small for this kind of analysis, about one fourth the minimum size specified by Linn and Harnisch. Recall that in the section of this paper on methods, Linn and Harnisch (1980) found that on groups with size about 500, standard errors were so large as to be of limited use.) The parameters from other groups were equated to the parameters from the white group. Then the following five indices were calculated as indicators of discrepancy:

1. Differences in discrimination parameters.
2. Differences in difficulty parameters.
3. Signed areas between the curves for two groups.
4. Unsigned areas between the curves for two groups.
5. A composite test of differences in both discrimination and difficulty parameters.

A one-parameter ICC model was also used to generate five similar indices. LOGIST was used again, with the specification that chance scores for all items were zero and discrimination values for all items were equal so that only the item difficulty value was estimated. These assumptions are those made by Rasch's one-parameter model. Differences in discrimination parameters are not tested using this model, since there are assumed to be no such differences. However, differences in difficulty parameters provide an index recommended by Draba (Note 35), signed areas between the curves for two groups form another index, unsigned areas form a third, and weighted differences in difficulty parameters using an asymptotic technique developed by Wingersky (Note 45) to account for the variance of the difficulty parameters form a fourth. The fifth procedure using the one-parameter model is based on how well the items fit the one-parameter model in the two groups. Within each group a fit statistic is computed as a mean square residual, i. e., a person's obtained response minus the probability of a correct response given the person's ability estimate. The absolute value of the differences in mean squared residuals for the two groups is the index of discrepancy.

The sixteenth procedure included in this study was the item discrimination method. Point biserial correlations were computed between items and total test scores within each cultural group. The index of discrepancy is the unsigned difference between those correlations for two groups.

To assure themselves that the data used in this study met the assumption of unidimensionality, Shepard and her colleagues factor analyzed the items for each Lorge-Thorndike test score separately on each ethnic group. In each case, using Reckase's (1979) criterion, they convinced themselves that the scores were unidimensional. (Their percentages of variance accounted for by the first factor did not quite reach Reckase's criterion of 20% in most cases.)

A special feature of this study is the inclusion of an external criterion for detecting bias. Previous studies have been criticized on the ground that the total test in which items are imbedded might be biased. As a result, biased items might not appear to be discrepant when evaluated against the total test score. Raven's Progressive Coloured Matrices test scores were used in this study as an external criterion. An important question is whether the same numbers of items and the same items will be chosen as discrepant using the external and the internal criteria.

The results of this study are about as confusing as the others. For six of the methods the number of significantly discrepant items were compared only to find that the numbers range from 30 to 63 depending on the method, the comparison groups, and whether the criterion is internal or external. Many more items are typically found to be discrepant when an external criterion is used, on the order of two to three times as many. For example, using the internal criterion, the signed C2 technique finds only 6 discrepant items on the verbal test when comparing Chicanos with whites, but finds 62 items discrepant for the same groups using the external criterion. On the other hand, using Wingersky's asymptotic method for measuring discrepancy with ICC's, exactly the same (large) numbers of discrepant items are found on the nonverbal Lorge-Thorndike score whether the internal or external criterion is used. Such results do not give one confidence that he can choose a procedure from this set to use in evaluating bias in a practical setting.

Rank-order correlations among the discrepancy indicators range from about -.70 to +.99. Of course, some of the methods are minor modifications of each other and would be expected to be highly correlated. Others are quite different in conception from each other. For example, the ICC one-parameter goodness of fit index of Wright, Mead, and Draba is very different from the others and turns out to have negligible correlations (-.39 to .42) with any of the other indices. Noticeably high relationships were found between the TID method and the methods based on one-parameter ICC differences in difficulty values and differences in signed areas. These data suggest that the simpler TID method could be substituted for the more complex methods based on the Rasch model quite acceptably. Some of the signed indices had substantial correlations with each other--on the order of .50--but the unsigned indices correlated neither with each other nor with the signed indices. The correlations among methods that can be compared with the Ironson and Subkoviak study seem to be in the same range, which is not reassuringly high. This consistency in modest values of correlation suggests that instead of expecting to find agreement among discrepancy indices and choosing the "best" one, the practical user will have to decide what kind or kinds of discrepancy he wishes to detect and use the index or indices that are sensitive to those kinds of discrepancy.

Shepard and her colleagues factor analyzed the scores from their indices. The four factor analyses based on comparisons between whites and the two other groups using internal and external criteria resulted in four regularly appearing factors all marked by ICC variables. One seemed to be an ICC one-parameter factor with signed values, another an ICC one-parameter factor with unsigned values, a third an ICC three-parameter signed factor, and the fourth an ICC three-parameter unsigned factor. It seems that using signed vs. unsigned indices and restricting bias to difficulty vs. letting it include discrimination tend to sort these various indices out to some extent. These may be four consistently different approaches to item discrepancy. However, some of the other methods did not load regularly on any of these factors, so their nature is not clarified in these data.

An attempt was made to identify discrepant items as outliers, i. e., items whose discrepancy indices were distinctly different from the remaining items. The authors' statement about some of the results seems fitting:

> Nevertheless, the agreements among methods, although modest, establish that the detecting of bias is not a random process. There is evidence of a consistent construct of bias upon which these methods converge. Still, it is disconcerting that in applied settings where only one method is used, the choice of method would make a substantial difference in the particular items considered to be biased. (p. 354)

For not a single discrepant item of the 90 items in the Lorge-Thorndike test do all the 16 methods agree in the comparison of blacks and whites using an internal criterion (the only data presented in a table). About half of the items are identified as discrepant by one method or another. Only three items are identified as discrepant by as many as ten of the methods. Only six items were discrepant according to at least half of the methods.

The authors suggest that while discrepancy-detection techniques should never be used as hard and fast criteria for rejecting test items, some methods are close enough to others that not all need to be used all the time. They suggest that chi-square could substitute for the three parameter ICC methods that consider only differences in difficulty or that consider signed areas between curves. The TID method is a good substitute for the one-parameter ICC methods that consider differences in difficulties or signed areas.

Scheuneman (Note 46). In the fourth study, Scheuneman (Note 46) compared three methods, the TID method, the C2 method, and a one-parameter method which compares difficulty indices (Draba, Note 35). The data were from a certification examination consisting of 175 items in 8 content areas. Each content area consisted of 12 to 33 items. The interesting thing about this study was that there were two forms of the test, and 75 items were common to both forms. The tests were administered six months apart, and only about 10% of the sample took both forms, permitting an evaluation of whether the same items would be identified as discrepant twice when they are mixed in with other items on the second administration, and when different subjects take the examination during the two administrations.

The subjects were 1350 whites and 115 blacks for the first administration and 1600 whites and 132 blacks on the second.

The TID method was used in a different way; the plot was made over all

175 items at the same time for each form of the test, ignoring problems of unidimensionality.

Scheuneman's C2 technique was used in a slightly different way also. Only three intervals were used, instead of the usual five. Also, instead of using all of the whites, only samples were used, 272 for the first test and 321 for the second. The probability associated with a chi-square of a given size was used as the index of discrepancy. (It was recognized only later that these probabilities have no meaning since the C2 index is not distributed as chi-square.)

Draba's index of discrepancy as the difference in difficulty level for an item when analyzed according to the one-parameter ICC model was used as a fourth index.

Correlations between discrepancy indices across the two situations are modest, .66, .63, .44, and 0.64 for the TID, C2, associated probability, and Draba methods, respectively. These levels of correlation for the same methods applied to the same items are not very encouraging. Scheuneman also examined the extent to which the same items were classified as discrepant on the two occasions. She selected the 20 most discrepant items on each administration and found that the TID procedure identified 9 of these items each time, and the C2 and Draba procedures identified 10 of these items each time. This is disappointing, too, though it is significantly better than chance. Scheuneman thought she might improve things by using two procedures at once to identify discrepant items, but this was not very productive, nor was the use of three at once. Scheuneman argued that the cost of misclassifying items is not large if the item pool is large. However, she forgot that in other reports she recommended deleting whole categories of items that tend to be discrepant, as we shall see in the next section of this paper. If whole topics are deleted based on erratic results from discrepancy indices, serious problems may result. For instance, the whole set of items involving negative language structures was dropped from the Metropolitan Readiness Tests because the C2 technique found many items of that kind to be discrepant (Scheuneman, 1979, 1980). A different technique or a different test administration might have found fewer of these items to be discrepant but might have caused some other part of the content to be deleted. One wonders how the content validity of the final form is to be justified.

Subkoviak, Mack, & Ironson (Note 47). Subkoviak, Mack, and Ironson (Note 47) compared the three-parameter ICC model, using areas between the ICC's as the measure of discrepancy, with the TID method, using perpendicular distance from the major axis as the measure of discrepancy, and the C2 method, as well as a formal chi-square statistic. The ethnic group's abilities were not adjusted to be equal, and they came from different geographic locations, 1008 blacks from a university in a large eastern city and 1021 white college students from a medium-sized midwestern city. The novel feature of this study was that items intentionally designed to be biased were inserted into a set of items taken from a widely used standardized test, the College Qualification Test (CQT). The intentionally biased items were taken from the Black Intelligence Test of Cultural Homogeniety (BITCH) developed by Williams using items from the slang of a black culture. Ten of these slang items were inserted among 40 items from the CQT. Within each block of five items on the test, one biased item was inserted. All of the items were four-choice multiple-choice vocabulary items. The subjects were instructed that some of the words were standard English, while others were slang.

Bias was coded as 0 or 1 depending on whether the item was from CQT or from BITCH. Correlations were calculated between the bias dichotomy and the level of discrepancy indicated by the four different discrepancy-detection methods. The ICC method resulted in a correlation of .87. The other correlations were quite similar; TID .73, C2 .71, chi square .71. ICC correlated with the other methods about .85; the other methods had intercorrelations about .95.

Even though the ICC method was used on samples that might be regarded as too small, 2000 in each group usually being recommended as the minimum, that method appeared to be substantially more valid in detecting the intentionally biased items. However, that is the method most likely to allow for differences in ability of the majority and minority groups. It is regrettable that the TID method used here was not based on ability-equated groups to see what its full potential might be. It is also regrettable that the authors did not report the degree of deviation from the major axis associated with the intentionally-biased items compared to nonbiased items. It has been suggested by Angoff (Note 16) that the criterion for regarding items as discrepant and potentially biased could be set at the level found in the most discrepant items appearing when extremely different groups from a common culture are examined. Another approach would be to set the level at the smallest discrepancy found between different cultures on items known to be biased.

Another problem with this particular study, and perhaps with other studies using intentionally-biased items, is that items that may have been designed to be intentionally biased for a particular black culture may not be biased for another black culture. The slang BITCH uses may not be recognizable in another ghetto, and the students used in this study were from cities widely separated geographically. (Angoff and Ford, 1973, found, for instance, that comparing blacks in Atlanta with blacks in Savannah resulted in discrepant items.) In this study we were were not informed whether the blacks were from the same ghetto culture used in developing the intentionally-biased items. If they were not, then the results might have been even more striking and the bias-detection methods more effective using more appropriate intentionally-biased items.

A related problem concerns the fact that the BITCH items are readily recognized by most whites as being biased. Such items are unlikely to survive preliminary screening and unlikely to be allowed to appear in a modern test distributed by a major test publisher. These are not the kinds of items that bias-detection methods are meant to locate--rather, those methods are meant to locate the biased items that have escaped careful scrutiny for bias by experts. It would be discouraging if the detection methods did not also locate items as baldly biased as the BITCH items, but the fact that a method does flag those items does not assure us that it will reliably and validly flag more subtly biased items without also flagging items which are unbiased.

Douglass (Note 48). There are two widely used computer procedures for applying the one-parameter item response theory model, BICAL (Wright, Mead, and Bell, Note 49) and LOGIST (Wood, Wingersky, and Lord, Note 50). Douglass (Note 48) points out that they are not identical, but maximize different functions because they handle items not reached differently. BICAL gives credit for an item only if it is answered correctly. Items omitted or not reached by the examinee are therefore counted as wrong. LOGIST considers only the items reached by the examinee, and for omitted items it gives credit as though the examinee had guessed randomly at the correct response. Items beyond the last one answered by the examinee are not considered in estimating the examinee's ability. Douglass shows that this results in a significant relationship between the amount of discrepancy in an item and its location in the test when using the BICAL but not when using the LOGIST one-parameter model. The test used was the 72-item Michigan State University Vocabulary Placement Test which is required for all entering freshmen. The subjects were 143 black males, 289 black females, 2954 white males, 3,121 white females, and a group of the 150 white females who were academically strongest. The test differed in speededness for these groups. Eighty percent of the blacks reached about item 50, 80% of the whites reached about item 65, and 80% of the talented white females reached item 72 (completed the test). LOGIST was used in a manner that resulted in its item difficulty parameters being scaled the same as the BICAL parameters.

The difference between the difficulty values obtained from two groups of examinees using the same computer program is the measure of item discrepancy. These differences were correlated with the positions of the items in the test. Positive correlations indicate that items with greater differences in difficulty parameters in the groups being compared occurred later in the test. The question is whether BICAL estimates or LOGIST estimates are least affected by speededness. If one procedure is less affected than the other, then it should be preferred for evaluating potential bias in speeded tests.

The results of this analysis were that the estimates using BICAL had significant correlations (from .30 to .56) between size of discrepancy and item position for all contrasts between blacks and whites, and a significant correlation of .37 for the contrast of white females with white academically talented females. The correlations from LOGIST ranged from - the .02 level. Clearly, BICAL estimates are influenced by test speededness.

Douglass also correlated the BICAL based estimates of discrepancy with the tests of goodness of fit provided by the BICAL program. These ranged between -.13 and .30, with only one of the 20 correlations involving blacks and whites reaching significance at the .02 level. Douglass concludes that the Rasch tests of goodness of fit cannot be expected to help in locating discrepant items.

We have considered above a series of studies which have used ICC methods and other methods comparing their results. While Linn and Harnisch (1981) did not review all of these studies, their summary of comparisons of ICC methods with others is not substantially altered by the additional studies. Their comment was the following:

> "While some agreement among the techniques was found, it is clear that the earlier techniques do not yield results equivalent to those based on IRT [item response theory] models which involve more than one parameter." (page 110)

Frary & Giles (Note 51.) In an empirical comparison of methods using live data, Frary and Giles (Note 51) applied their method of using person-fit statistics to detect item bias to the tests of Virginia's statewide testing program and the Virginia Graduation Competency Tests. The statewide testing program involves eighth graders and includes tests of the following: Reading Vocabulary, Comprehension, Reading Essential Skills, Study Skills, Math Computation, Math Problem-solving, Math Essential Skills, and Occupational Information. The Virginia Graduation Competency Tests measure just reading and mathematics. The subjects for their study were samples of 300-400 cases from those who omitted less than 10% of the items. Four different person-fit statistics were obtained for each test and averaged across tests. The four statistics were the one-parameter ICC person fit statistic, the Frary-Giles statistic measuring how well each person's choice of an alternative fits his expected choice summed over items, a similar measure summed over items after standardization across items, and Good's (1956) measure of the unusualness of an examinee's response. The last three methods tend to have intercorrelations near .90 for each test. The one-parameter ICC fit statistic has appreciably lower correlations with the other three methods, ranging from .27 to .76, the highest values being with the person fit measure standardized before summing across items.

Frary and Giles attempted to see whether person-fit statistics varied across the variables of sex, race, size of community, and level of test scores. Random samples of combinations of these variables were drawn, and a series of analyses of variance were computed. Evidence of sex, race, and community-size significant effects was scattered. The most pronounced and consistent evidence of both sex and race effects was found for the low scorers, with whites showing fewer unexpected responses than blacks and females fewer than males. Attempts to locate the items that contributed most to person-fit differences resulted in the conclusion that there were relatively few such items except on the Reading Essential Skills test of the statewide testing program. There was no obvious explanation why any of the items contributed strongly to the person-fit differences. The results are not reasonably interpretable in terms of socially or culturally induced bias.

(PLease continue on next page.)

Applications of bias detection methods in practical situations

Chi square

Scheuneman has perhaps been the most audacious in using the results of item bias detection techniques to modify tests that are to be widely used in practical applications. In two publications (1979, 1980) she reported use of the techniques during test development of the Metropolitan Readiness Test. Over 500 items from 44 subtests in 7 skill categories included in the final tests were analyzed for difficulty discrepancy. The C2 technique using an alpha level (appropriate for chi-square of the same numerical value) of .30 indicated discrepancy in 76 items, about 14% of the total. Most of the discrepant items were in the language area, and in that area many of the discrepant items involved negative language structures. Therefore, all items that included negative language forms were dropped from the item pool. Later, during standardization when it was no longer possible to modify the tests, large groups of minority children became available. Discrepancy analyses then found that items involving artificial letters favored Orientals, and items involving comparatives handicapped blacks. Items involving complex grammatical forms favored black and white children over Orientals, Puerto Ricans, and Mexican Americans.

Scheuneman (1980) briefly reported similar analyses of the Otis-Lennon School Ability Test. There, 30% of the verbal analogy items at the junior high school level and 50% at the senior high school level were found to be discrepant. Letter- and figure-series items were also discrepant. It is apparent from Scheuneman's writing that the procedure she regards as proper is to eliminate from consideration the item pools that produce these large numbers of discrepant items. Thus the items involving negative language structures were removed from the Metropolitan, and if it had been discovered in time, presumably artificial letter items and letter- and figure-series items would no longer be used. This raises some interesting questions. How does one discover that students have difficulty with these kinds of content, especially negatives, comparatives, and complex grammatical structures, if they are not tested? Should we no longer expect children to learn to understand written negatives, comparatives, and complex grammatical structures? Furthermore, one might ask whether such items would have been deleted if some of the other bias-detection techniques had also been employed and had not found these kind of items to be discrepant? (Would these have been found to be discrepant if chi-square with a sound test of statistical significance had been used instead of C2?) Would the kinds of items detected as discrepant by each of the other techniques also be eliminated from the item pools? What kinds of items would remain? What would the process of item elimination have done to the content validity of an achievement test?

Item discrimination

Green used the item discrimination method in the development of the 1970 edition of the California Achievement Tests (Green, 1972). The tests involve reading, mathematics, and language skills. Students were obtained from seven groups of schools to represent the following:

a. Northern, black, central city.
b. Northern, white suburban.
c. Southern, black rural.
d. Southern white, suburban.
e. Southwestern, Mexican American.
f. Southwestern, Anglo-American, suburban.
g. Southern, white rural.

The grade levels were 1, 3, 5, 8, and 10, and sample sizes ranged from 144 to 361. A total of 274 separate analyses were conducted.

It will be recalled that the purpose of the method was to select items on the basis of item-test correlations within each of the seven groups on each test in each grade. The best half of the items in terms of item-test correlation were chosen. Then items which were chosen for one group but not the other were regarded as discrepant. Scores for each group in a contrasting pair were obtained on both sets of discrepant items. The correlation between scores on the two tests was found. Reliabilities of the full-length tests and the tests based on the best half of the items were compared, and mean scores on the full test, the best half tests, and the discrepant items were computed.

The proportion of items classed as discrepant varied from .15 to .70. The only pattern seemed to be that more were discrepant for Mexicans compared to Americans than for blacks compared to whites. The median correlation between tests using discrepant items is about .5, so tests developed on the basis of different cultural groups may measure different things. Little bias was indicated by the reliability comparison. The items chosen on the basis of the majority were almost uniformly more difficult than those chosen on the basis of the minority. Green concludes that using this procedure and definition of bias, most tests are probably biased against most groups. However, he observes that the amount of relative improvement that a minority group could expect by using test built with items selected on tryout groups like itself would be small and unlikely to overcome any large difference between group means. Instead of deleting the content that appears to be biased, Green suggests that test constructors build alternate test versions, each based on an item analysis for a particular group. He regards the difficulties involved in doing this as large but not insurmountable.

TID

Arneklev (Note 52). Jensen (1980a) described a study by Arneklev (Note 52) using data from the five subtests of the Comprehensive Tests of Basic Skills. The tests include the following topics: Reading Vocabulary, Reading Comprehension, Arithmetic Computation, Arithmetic Concepts, and Arithmetic Applications. The subjects were white, black, and Asian pupils, but Jensen described only the black vs. white comparisons since there were relatively few Asians. Arneklev used several different item-discrepancy indicators, but since the results were similar, Jensen described only the analyses based on item difficulties. Items were considered discrepant if they deviated significantly from the regression line (not the major axis) of white p values on black p values. Arneklev also correlated p values for the items, a closely related procedure. The cross race correlations of p values were very similar to the within-race cross-sex correlations; all of them were very high. In the TID analyses, only 6 to 9 items were found to be discrepant in any contrast between sexes or ethnic groups. Out of the total of fifteen different items that were discrepant in cross-racial comparisons for males and females, 12 were easier for blacks and three were easier for whites. Eliminating the discrepant items would have the tendency to increase the mean white-black difference in total scores.

Raju (Note 53). Raju reported to the Third Annual Johns Hopkins University Symposium on Educational Research in 1980 the work that Science Research Associates has been doing on the problem of item bias in the SRA Achievement Series and the Educational Ability Series tests. The procedure used on all forms and all levels of these tests was basically the TID approach. Both at pretest time and after standardization, data were gathered from white students and from black, Hispanic, an American Indian students. Item difficulty values (p values) were obtained for all items separately for each form and level and each cultural group. At pretest time, the items that did not meet traditional criteria for satisfactory items in terms of difficulty and discrimination were rejected, and the rest were analyzed for sex and race discrepancies, male vs. female and white vs. each of the other ethnic groups. The p values were converted to the delta scale, and then they were plotted on two-dimensional plots. For some levels, the next step was simply to handdraw an ellipse around the points so as to include the greatest number of items within the narrowest possible ellipse. Items lying outside that ellipse were regarded as discrepant. For Form H, the major axis was drawn through the plot, and a distance from that axis was calculated for each item. (Raju did not state why the procedure was changed for Form H or what level of deviation from the major axis was used to designate items as discrepant.) The only general results reported for the use of the procedure were that in the pretest there were higher percentages of items deviating far from the major axis of the ellipse for the black vs. white comparison than for the male vs. female comparison, and for the reading test than the mathematics computations test. Comparing the pretest results with the results after standardization shows that the deviations from the major axis were smaller for the standardization data.

Bode (Note 54). In a paper presented at the annual meeting of the American Educational Research Association, Bode reported additional studies of the Form 1, Level E data for the 1978 edition of the SRA Achievement Series. The sample sizes for several of the groups were identical to those reported by Raju, so it may be that the same subjects were used for both of these analyses. In Bode's studies, the pretest analyses, using a TID method with handdrawn ellipses described just as Raju described his, were compared with results from later analyses to see whether the same items appeared to be discrepant. The groups contrasted were males vs. females, whites vs. blacks, and whites vs. Hispanics. It appears that nearly 2000 items were pretested altogether, but only about 300 were discrepant, so the rate of discrepant items is quite low compared to results from some other researchers. Of course, many of the discrepant items from the pretest data were eliminated from the test. Twenty-five discrepant items were retained for unspecified reasons.

The retests occurred 12 and 18 months after the pretest. The retest data came from the standardization groups. There may have been no overlap in examinees with the pretest group. The minority groups for pretest were identified by their schools, but in the standardization testing examinees described their own ethnic groups. Sexes were compared within ethnic group on the pretest but across ethnic groups on the standardization. In the standardization data, discrepant items were identified by the TID method based on distance from the major axis. Discrepancies greater than 1.0 indicated bias. (Bode does not say how that discrepancy was calculated, but presumably the procedure used by Angoff and Ford, 1973, Angoff and Herring, 1976, and Angoff and Sharon, Note 55, was used.)

In order to compare pretest results with standardization results, the pretest data were reanalyzed using the major axis and discrepancies greater than 1.0. Using this technique, the number of discrepant items retained for the final test form increased from 25 to 43. In many cases different items were found to be discrepant using this method instead of handdrawn ellipses.

When Bode examined the discrepant pretest items to determine whether they were also evaluated as discrepant in the standardization analyses, she found little overlap. Only one item was identified as discrepant in all three analyses. This disagreement among analyses could be attributed to a number of possible factors. First, of course, is the possibility that the technique does not identify discrepant items soundly. Second, that which is discrepant at one time for a group of students is not necessarily discrepant 12 or 18 months later. For example, an achievement test item may be unknown to some students in a group before instruction, or known only to highly-motivated students or to students from an advantaged background. After all of the students have been instructed on the topic, the content may be understood thoroughly by all students, and the item may no longer be discrepant. Third, what is discrepant in the context of pretesting is not necessarily discrepant in the context of standardization testing. Fourth, sample sizes were too small to reveal large deviations from the major axis consistently. Bode says that even though one standardization sample consisted of 129,000 students in 457 schools in 92 districts from 38 states, it only contained 100 Hispanics. However, more than two-thirds of the total number of items identified as discrepant came from the white vs. Hispanic comparison. Other differences between the pretest data and the standardization data might be responsible for the failure of cross validation. A tighter design would be highly desirable.

Bode found that discrepancies occurred in both directions, some items relatively easier for the majority group and some favoring the minority group. The more difficult items tended to be even more difficult for the minority group and the easier items tended to be more difficult for the majority group. When items were deviant toward the Hispanic sample, the white sample had about twice the percent of omits of the Hispanic sample. Complex interactions between discrepancy and the proportion not reaching an item were also observed. These analyses again suggest that some of what has been called item bias may be a matter of guessing behavior or item location in a speeded test, instead of item content.

Coffman (1978). Coffman reported analyses of subtests of the *Iowa Tests of Basic Skills*, levels 12, 13, and 14. They had been administered to samples of black and white students in a small city. Roughly 200 students of each race were tested in each of grades 6, 7, 8, and 9. The differences between the item *p* values for the two races were compared after transforming the *p* values with the arcsin transformation. Coffman noticed that the sizes of differences seemed greatest on subtests in which development of the ability depended more heavily on outside-of-school factors. Like Bode, he noticed that items that were too difficult for the examinees tended to favor the blacks. Also like Bode, Coffman had several samples and checked to see whether the same items were discrepant in different samples. Out of a total of 795 different items in the study, 109 (14%) were outliers in two or more samples. Of these, 35 were very difficult items so Coffman suggests that random responses may account for the discrepancies in *p* values. Of the remaining 74, about the same number were relatively easier for blacks as were relatively easier for whites. That is, the "bias" was balanced.

Angoff & Sharon (Note 55). One of the early uses of the TID method was that by Angoff and Sharon (Note 55) in 1971. They were concerned about differences among foreign students in performance on the various sections of the *Test of English as a Foreign Language (TOEFL)*, and they also evaluated differences among them on items within each section. ANOVA revealed significant and important section by group interactions and item-by-group interactions. The TID method was used by comparing each group with every other group, but that was too many plots for an article, so they also compared each foreign group with a reference group drawn at random from the entire candidate population of about 14,000 who took this test in October of 1969. In addition to TID plots, Angoff and Sharon calculated the correlations represented by the plots, suggesting that linguists might find most useful and interesting the plots with the lowest correlations, e. g., r=.689 for the Germans vs. the reference group. Angoff and Sharon made no attempt to determine the cause of item-by-group interactions or to examine aberrant items for consistencies among kinds of items causing interactions.

Angoff _&_ _Herring_ _(1976)._ The TID method was used with a majority group comparison as the baseline in this 1971 study comparing the performance of Canadian and American students examined by the _Law_ _School_ _Admission_ _Test_ (LSAT). Two samples of American examinees were used, both of whom took the same form of the LSAT but at different administrations. About 1000 examinees were in each of those samples. The Canadian group of 500 students were all those who took the same form as the Americans, but were not from test centers likely to have large numbers of French-speaking Canadians. Delta plots were formed after ascertaining that the groups were of very similar ability levels. The two American samples were plotted against each other to obtain a base line for judging items as discrepant in the plots contrasting Americans with Canadians. A 95% band was drawn around the principal diagonal of the correlation ellipse in the American vs. Canadian plots. The outlier items were examined, and nothing systematic could be observed about them. Further, the items that were outliers in one plot of American vs. Canadians were not outliers in the other plot comparing those two cultural groups. Correlations between deltas were computed and ranged from .95 to .99. Angoff and Herring concluded that the items were as appropriate for these Canadian examinees as for American examinees.

Wightman (1976). Wightman repeated the Angoff and Herring study on a new form of the LSAT after some changes had been made in that test. He used about 2000 Canadian examinees and about the same number of non-Canadians (predominantly examinees at United States testing centers). Delta plots were examined comparing Canadians with non-Canadians, and, with no discussion of significance tests or other techniques for detecting which items were discrepant, it was concluded that all points plotted were close enough to the line of relationship to justify saying that there was no cultural bias in the items. Looking at the plots, the reviewer would have to agree with that conclusion.

Donlon (Note 56). Content factors associated with sex differences in test questions were studied in connection with the College Board's SAT administered in 1964. Donlon used results from about 50,000 boys and a similar number of girls tested in May of that year. With such large numbers of cases, only rarely was there not a significant difference between the p values of the two sexes for an item. Donlon apparently did not construct plots, or transform p values to deltas. Instead, he decided that differences of .07 in p values were large enough to be of interest. Using that criterion, he found 19 items on SAT-V which favored one sex or the other, 8 favoring men and 11 favoring women. He felt that the content difference between these items was the same one reported by Coffman (1961) on SAT-V ten years earlier, i. e., men were favored by mechanical or business content and women by personal or personality content. In terms of item types, antonyms tended to favor women and analogies to favor men. Reading comprehension favored women by the largest amount, and Donlon concluded that the reading comprehension items were the source of women's superior performance on SAT-V.

On SAT-M, the items generally favor men. On only two items did women do better. Seventeen of the 60 items are judged to have "male flavor," that is, to involve mechanical relationships, refer to real world things, have male characters performing, and so on. The women do their best on straight algebra items. If the content of the entire test were limited to algebra, the score difference on the total test favoring men would be halved.

Donlon, Hicks, & Wallmark (1980). Strassberg-Rosenberg and Donlon (Note 3) analyzed the SAT of April 1974, ten years later than the data of the Donlon study reported immediately above. Their outlier analysis tended to confirm the results of the two previous studies. This result created interest in analysis of the items of the Graduate Record Examination Aptitude Test (GREAT) which has different specifications than the SAT, is taken by an older group of examinees, but also is parallel to the SAT in many ways. Donlon, Hicks, and Wallmark examined data from the GREAT administration of December, 1974, using nearly 2000 white males and a like number of white females. Delta values were calculated for each item for each group. These were plotted, the major axis of the ellipse was determined, the perpendicular distances from that axis were obtained, the standard deviation of those distances was calculated, and items whose points were more than 1.5 s. d. units from the major axis were considered to be outliers. Among the 55 items on the vocabulary test, 5 were outliers. Again the content of the 2 items favoring males tended to be scientific and practical, and that favoring women tended to be related to human affairs, though the data supporting this conclusion are very limited.

Analysis of deviancy according to kind of item yielded results (Donlon et al., 1980, p. 14) "directly contrary to the patterns in the SAT analysis by Strassberg-Rosenberg and Donlon" (Note 13). In the GREAT data, the antonym items favored females, while the analogy and sentence completion items favored males. Donlon and his colleagues say that this casts doubt on the conclusion reached earlier that sex biases were related to item types. Among the 40 reading comprehension items, there were six outliers. The outliers seemed to be somewhat passage dependent--five of the six were accounted for by two of the passages. However, not all items on these passages were outliers. In the 55 mathematics items, nine were outliers, five favoring females and four favoring males. Those favoring males were related to a graph concerning changes in personal income. Those favoring females involved dealing with algebraic unknowns, a phenomenon noted in Donlon's earlier study. Since there were few items in the GREAT that had reference to persons, no clear data on Donlon's and Coffman's earlier "male world" hypothesis were available. Donlon, Hicks, and Wallmark noted that the females' scores on the math section could be increased by replacing items based on graphs by items based on algebra, but they regarded this as questionable since the ability to handle graphs is a legitimate part of quantitative skill.

Swineford (1976). The December 1971 administration of the LSAT was studied for item difficulty discrepancy between whites vs. blacks and whites vs. Chicanos. The sexes were separated, so three contrasts were made, white males vs. black males and vs. Chicano males, and white females vs. black females. Swineford did not report the numbers of Chicano or white subjects, but reported the number of black males as about 600. She reported that the groups differed in ability, but adjustments were not made for those differences in the analysis. Although it was not explicitly stated, it can be deduced that the delta plot methodology often used at Educational Testing Service was used in this analysis. Eleven items were easier for white groups than for black and/or Chicano groups. None was easier for whites in comparison with more than one subgroup. Twenty items altogether were judged to be outliers, but only one appeared in all three comparisons, and it was easier for the minority groups. Nine items were easier for the minority groups, and five of them referred to a single reading passage dealing with social customs of certain Indian tribes. Swineford concluded that the results provide no basis for suspicion that some items are biased relative to other items of the same type.

Sinnott (Note 5). The Graduate Management Admissions Test (GMAT) was studied by Sinnott. She did five analyses, one contrasting sexes, another contrasting races (white, Oriental, Hispanic, and black), the third considering race and sex, the fourth, age, and the fifth contrasting with whites seven foreign candidate groups proficient in one of the following languages: Chinese, English, French, Indo-Iranian, Japanese, and Spanish. At least 400 candidates were in each minority group; a minority group was always contrasted with the majority group rather than with another minority group. These examinees were sampled from those who took the GMAT in January, 1978. The usual Educational Testing Service delta values were calculated for items, and the major axis of the ellipse formed by plotting pairs of deltas for a contrast was calculated. However, deviant items were then removed from the set, and a revised axis was located based on "nondeviant" items. This process was repeated until the same items were deviant on two successive analyses. Then the items which had been removed for fitting of the line were reintroduced to determine their distance from the final line. Empirical comparisons within cultural groups were used to help determine what cutoffs would be used to regard items as outliers in the analyses. For the first four analyses, a delta distance of .7 was the cutoff. For the fifth, a delta distance of 1 was the criterion.

Analyses were done separately for the six sections of the test. In reading comprehension, analyses typically yielded only one outlier item. In problem solving and data sufficiency for some analyses there were relatively few deviant items, but for others deviant items were more common. The age group analyses had no deviant items in some cases, while the language groups had as many as 8 or 10. The older examinees may have had more difficulty with formal mathematics items than with word problems involving daily affairs. The foreign students had unexpected difficulty with these real-world problems. In the practical business judgment section of the test, some analyses showed no deviant items. Others showed as many as 5 or 6. Sinnott suggests that the pattern might be due to some of the examinees' misunderstanding the item instructions. The data sufficiency section had some analyses with no deviant items and some with as many as a dozen. Again, the older examinees had difficulty with nonword problems, and the foreign language groups had difficulty with the word problems. Finally, in the language usage section, the number of deviant items again ranged from none in some analyses to as many as 8 in others. The items easy for foreign language groups were those on which they might had have extensive drill and practice. Those easier for the majority were concerned with subtle points. Nowhere in Sinnott's report is there an indication that outliers were due to irrelevant factors. The one possible exception might be the misunderstanding of the instructions for the practical business judgment items.

Carlton & Marco (Note 37). Carlton and Marco reported on a number of studies of item bias that have not been examined directly by the reviewer. Cowell and Swineford are reported as having studied the performance of white males and white females on the LSAT items. Delta plot methods were used, and each section of the test was treated separately. Sample sizes were about 1200 of each sex. Within-sex plots were used as the guide for setting cutoff levels for deviant items. No deviant items were found.

Carlton and Marco describe briefly a study of the National Teacher Examinations comparing black and white students using TID methods. B. J. Humphry found seven of the 109 items to be outliers. She also examined occasions in which a particular incorrect option was more popular for blacks than the correct option, and she examined the blacks' patterns of omits. No details of findings in these analyses are reported by Carlton and Marco.

Carlton and Marco report that beginning in 1973 item bias has been studied on a number of different forms of the SAT and the Test of Standard Written English (TSWE). The studies have been conducted by Stern, by Cook and Stern, by Blew and Ishizuka, and by Blew and Stern. Each study used TID methods. One used the innovation of matching samples on ability. In each study the correlations of item difficulty levels across samples were reported for different parts of the tests. The lowest of these values was 0.908. In each analysis deviant items were located and examined for content clues as to the reason for the discrepancy. The number of discrepant items identified is on the order of a dozen for verbal and another dozen for mathematics in studies with unmatched samples and about half that many in the study with matched samples which also used a more stringent standard for deviancy. Typically 1000 examinees in each comparison group were used. Carlton and Marco did not report any results of the content analyses of the deviant items. They indicated that these studies of the SAT and TSWE will continue and that public reports will be made of the results.

Rudner (1978). Rudner (1978) studied discrepancies in item difficulties in order to guide curriculum developers and teachers of the hearing impaired as to concepts that require special attention. The test was the Stanford Achievement Test, Intermediate I Battery (1964 version) with the following parts: Mathematics Computation, Mathematics Applications, Mathematics Concepts, Spelling, Social Studies, Science, and the Language subtests of Word Meaning, Sentence Sense, Word Usage, Punctuation, Capitalization, and Dictionary Skills. The examinees were 2,821 hearing-impaired and 618 normal people. The index of discrepancy used was the TID method in which items that were ".75 z units" away from the major axis were classed as discrepant. Plots were made for each of the subtests, but data were only summarized in the article. The largest proportion of discrepant items was found in the Sentence Sense and Dictionary Skills tests and the smallest proportion in the Spelling, Mathematics Computation, and Capitalization tests. Twenty-two items were relatively easier for the hearing impaired, and 26 were relatively easier for the normal subjects. The percent of discrepant items ranged from 2 to 44. The hearing impaired seemed to have particular difficulty with negatives and with comparatives. Rudner, instead of suggesting that these concepts be removed from tests for the hearing impaired, suggested that the curriculum developers and teachers pay extra attention to them. This consideration of possible bias against people with various specific handicaps greatly enlarges the number of groups for which special tests would have to be developed if Green's suggestion is taken seriously.

nalysis of Variance

Sandoval (1979). Sandoval studied item bias in the WISC-R for children 5 to 11 years old. He used subjects from the sample used by Jane Mercer to standardize the System of Multiculture Pluralistic Assessment (SOMPA). The 1050 cases were approximately equally distributed across sex and three ethnic groups, whites, blacks, and Mexican Americans. Each subtest of the WISC-R was evaluated separately, and several methods were used, including rank order correlations of item difficulties and rank order correlations of differences in adjacent-item difficulties, methods described by Jensen (1980). However, the larger-scale analyses were ANOVA's as suggested by

Cleary and Hilton (1968), and a multivariate ANOVA which examined the main effects of ethnic group and socioeconomic status and the interaction of the two. The correlation analyses yielded high correlation values casting doubt on the existence of bias in the items. The item by ethnic group interaction terms were almost all statistically significant with samples as large as these, but the contributions to explained variance were all less than one percent except for Block Design test which reached 1.08%.

The multivariate analysis of variance indicated that the interactions between ethnicity and socioeconomic level were not significant, and that differences in performance for ethnic groups occurred on all subtests. Sandoval concluded that the notion that there may be a number of items with radically different difficulties for children from different ethnic groups was not supported. Of particular interest may be the fact that Sandoval noted that the single item most often quoted as unfair in WISC-R, the sixth item of the Comprehension test (boys fighting), showed no significant difference in difficulty between groups. No logical pattern was found among items that were harder for a minority group than for the white group.

Cotter and Berk (Note 57). Cotter and Berk also studied WISC-R, comparing 112 black, 126 white, and 117 Hispanic learning-disabled children between ages 10 and 10.33 from a city in Florida. The bilingual children in this study were evaluated by bilingual psychologists who translated WISC-R items as they saw fit. Unlike Sandoval, Cotter and Berk matched their subjects on sex and on subtest raw score, which greatly reduced their sample sizes. They had between 60 and 88 black-white pairs and between 56 and 98 Hispanic-white pairs for the nine WISC-R subtests they studied. They used both judgmental analyses and statistical analyses to locate discrepant items. Six black educators and three Hispanic educators reviewed all the items in WISC-R looking for subtests or items that were biased against their ethnic groups. In the statistical analysis, item p values were converted to the delta scale and then used in an item-by-group ANOVA separately for each subtest. Whites were compared with blacks and separately with Hispanics. Cotter and Berk attempted to take statistical power of their tests into consideration. They then also had judges review the items that were discrepant in the statistical analysis. In these judgmental analyses, nine judges were used in each group, including a panel of white judges.

The first panels, the ones who judged all the WISC-R items, concluded that essentially none of the subtests or items was biased against blacks, but 8 items were biased against Hispanics, all of them being in the Information and Comprehension subtests. The subsequent statistical analyses found significant item-by-group interactions at the .05 level for those two subtests for blacks vs. whites, and for those subtests plus the vocabulary subtest for whites vs. Hispanics. Followup analyses using the Newman-Keuls test indicated that some items favored the minority group and some favored the majority group. A total of 13 items were found to be discrepant in black vs. white analyses, and 7 in Hispanic vs. white analyses.

When the second set of committees reviewed the items that had been found to be discrepant in the followup statistical analyses, the black review committee disagreed with the statistical analyses on 11 of the 13 items, finding no reason for these 11 items to be biased in spite of the statistical results. The Hispanic review committee found reasons for the statistical discrepancy for 4 of the 7 items that were discrepant. The white review committee agreed with the statistical analyses on 3 out of 13 items, but they could not give a reason for one item and could not agree on a reason for another. Even when effect sizes are large, as for one item on the Similarities test, the committees could not give a reason for the discrepancy. In this study the often discussed item 6 on the Comprehension subtest, the item about children fighting, was deviant against whites instead of blacks as is often claimed (Williams, 1970,1971) although the effect size was small. In summary, Cotter and Berk note that the significant interactions among subtests account for less than 2% of the respective total subtest variances, suggesting that there is very little bias in the subtests. Further, a relatively low percentage of the items flagged by judges as biased were found to be discrepant by statistical analyses. This points again to the ineffectiveness of judgments by experts as a method of detecting bias that will also appear in statistical analyses.

Item difficulties

Flaugher and Schrader (Note 23). Flaugher and Schrader (Note 23) attempted to eliminate bias by eliminating items simply on the basis of differences in difficulty for black and white students. They did transform the difficulty values with a normal curve transformation, as in the TID methods, but they did not make a scatterplot of values. They identified as discrepant any items with transformed difficulty value differences greater than 3 on the delta scale. The test was the College Board's SAT. The subjects were 1000 blacks and 995 whites selected as spaced samples from those taking the test in December of 1974. Twelve items were identified as discrepant out of the 85 items in the SAT verbal section. Fifteen were identified as discrepant from the 60 items in the SAT mathematics section. These items were removed from the test, and the data were reanalyzed. Essentially, the removal of the discrepant items did little or nothing to equalize the performance of blacks and whites. The blacks' performance was about 1 s.d. below the whites' with all the items or with only the items of equal difficulty. There was a distinct positive relationship between size of the item-test correlation and differences in item difficulty, with large differences associated with highly discriminating items. The authors concluded that the effect of removing these discrepant items on differences in black and white performance is not much, a finding we noted earlier in a study reported by Bianchini. However, removing the discrepant items would make the test considerably harder since the items selected for elimination tended to be easier items. Making a hard test even harder does not seem to be a sound approach to test development since the original test was already quite difficult for blacks.

Subjective judgments of items

Sandoval and Miille (1980). Sandoval and Miille (1980) made specific tests of the questions of whether judges could detect items that were harder for minority than majority students, or vice versa, and whether the effectiveness of judges for this purpose depended on their ethnicity. They used items from the WISC-R that had been studied earlier (Sandoval, 1979) and found to be differentially difficult or equally difficult for whites and blacks or whites and Mexican Americans. Fifteen equally difficult and fifteen items of similar difficulty levels but distinctly favoring whites were assembled and evaluated by judges. The items were grouped according to subtest. The information for the judges included the administration procedures from the manual, the actual test items, and the corresponding answers, including information on partial credit. The judges were 100 students at the University of California, Davis. Thirty-eight of them were self-identified blacks, 22 were Mexican Americans, and 40 were whites. All were undergraduate students. They were asked to decide whether each question was more difficult for whites, for blacks or Mexican Americans, or was equally difficult for both groups being compared. Separate questionnaires were used for the white vs. black comparison and for the white vs. Mexican American comparison.

The results indicate that judges are not able to differentiate effectively between the differentially difficult and equally difficult items. Overall, they were able to identify the relative difficulty of the items only 45% of the time. By chance one would expect 50% correct identifications. All judges tended to assign about 70% of the items to equal difficulty. However, for the items that favored minorities, the judges claimed equality about 75% of the time, and for those that were equally difficult for both groups, the judges only claimed equality about 65% of the time. These results are the opposite of what would be expected if the judges were accurate. The whites responded with the same level of accuracy (or inaccuracy) as the Mexican Americans or the blacks. There appeared to be no pattern among the questions that were found to be more difficult. Sandoval and Miille (1980, p. 252) conclude that "these results refute the claim that subjective judgments can be used to determine item bias without the aid of empirical evidence (Williams, 1971)."

Summary

The issue of item bias arises from differences between the mean scores of different cultural groups on sets of items or tests. The rationale for invoking the concept of item bias rests on the assumption that the groups are actually equal in performance on the underlying construct being measured but that an unfortunate (or deliberate) choice of items has caused the appearance of inequality in performance. It is reasoned that even though experts can see no bias in looking at the items, still enough bias can have crept into some or all of the items that statistical procedures can ferret out that bias.

Since there could be real differences between cultural groups in test performance, just as there are differences in other characteristics about which there can be no reasonable dispute, one must define item bias carefully. A difference between group means does not automatically signify bias in items. It appears that item bias is best defined as a difference between groups in performance on an item that is greater than usual among items on the same unidimensional measure and that is due to irrelevant aspects of the item. Investigators do not agree completely whether item bias is best treated as a dichotomy (requiring definition of "unusual") or as a continuum in which all items are biased but some are more biased, and one seeks to reduce bias to a level at which it is of no practical importance. One seeks to reduce item bias in order to be sure that score differences are not due to artifactual effects. In some cases, one may try to locate biased items so that any bias in mean test scores of different groups, such as males and females, or students who have studied different texts for the same course, can be cancelled out by including equal numbers of items biased in favor of each such group. It is conceivable that biased items might mask true differences between groups. Such phenomena would go unnoticed without analysis to locate biased items.

Many methods have been proposed and explored to evaluate bias in items. Judges can note potential for bias when it is glaring, but they may not be effective in noting bias when the bias is subtle. Statistical methods look for discrepancies between performances of different cultural groups on items. Discrepant items may not be biased; the cause of the difference may be relevant, or no cause may be discernible. In these cases the definition of biased item is not met because the definition includes the specification that the bias be due to an irrelevant aspect. Items which the statistical methods detect as showing discrepant performance by different cultural groups must then be evaluated by expert judges to determine whether the cause of the discrepancy can be located, and if located, whether the cause is indeed irrelevant.

It would be reassuring if the many methods agreed generally in deciding which items were potentially biased, and if they generally chose items as potentially biased which were found by judges to be biased in the sense that the causes of the group differences could be located and the causes were irrelevant. A large body of research has compared the various statistical methods to see whether they tend to pick out the same items as biased. Much less research has concerned itself seriously with locating causes of bias and evaluating their relevancy. We have described and discussed 9 categories of methods:

1. Comparisons of item difficulties
2. ANOVA
3. Chi-square
4. Correlation
5. Item-response-theory
6. Comparisons of foils or distracters
7. Factor analysis
8. Experimental manipulations
9. Construction

The research on item bias seems to have started in about 1960, and since then studies have evaluated single methods on real data, compared methods on simulated data, compared methods on real data, and applied the methods in practical settings. While there is enough agreement among methods in detecting items which might be biased so that one could conclude that biased items could be created or could be found in tests and the methods would be helpful in locating them, there is so little agreement among the methods that in practical use which items and how many will be judged potentially biased depends heavily on which method is used. A different method could yield strikingly different results.

No method is without its limitations. Nearly all of them share the assumption that the characteristic being measured by a set of items or a test is unidimensional. Most methods assume that the ability or achievement levels of the groups are equal. Many of the studies and applications of the methods have not tested or met these assumptions. Some studies have been done on groups of equal ability or have attempted to create groups of equal ability by selecting from one group subjects similar in ability to the other group. Usually when groups are equal in ability or equalized, few items are notably discrepant. However, only items that are discrepant with equal ability groups could be the cause of mean differences between groups of equal ability, the issue which raised the problem of item bias in the first place.

The idea of "notably discrepant" is a troublesome one. For those who argue that there is a dichotomy between biased and unbiased items, and the biased items are outliers in a bivariate plot, there must be a definition of outlier. No clear definition has been established. Statistical tests do not provide a clearly satisfactory solution. With large samples of subjects, items that are quite similar in difficulty level will be detected as potentially biased using the usual statistical significance levels. However, these differences may be so small that they have no practical significance, and obviously, the number of potentially biased items depends on the size of the sample, a situation that is not very satisfactory. Furthermore, even when a number of items have difficulties that are quite different for the groups being compared, removing those items and replacing them with other items drawn at random from the same pool of items (or even removing them without replacement) does not appear to remove the differences in mean scores for the groups being compared. The practical outcome that is desired by those who attack a specific test on the grounds that it contains biased items is not achieved. In fact, it seems unlikely that large differences in means between groups can be removed by replacing discrepant items with other items sampled from the same item pool. Small differences, such as those sometimes found between sexes or between students of the same topic but using different curricula may sometimes be reduced either by replacing discrepant items or by balancing discrepant items so that some are found easier by one group while others are found easier by the other group.

In the second section of this paper which discussed problems in detecting biased items, a set of five important conclusions from the research on item bias were listed. Let us now consider the support for these conclusions.

First, we concluded that judgment by people versed in detecting bias in items can be relied on to remove severely biased items. However, such judgments may err in removing items that are not biased, thereby removing from the available pool items (perhaps many) which might be perfectly useful. Examples of situations in which items were found to be biased by statistical analysis, but in which judges could readily have discerned the bias, and, indeed, test developers would probably not have used the biased items in the first place if they had been sensitized to the issue of bias, are the study by Durovic (described by Scheuneman in Note 6) and by Subkoviak, Mack, and Ironson (Note 47). In the former, Durovic is reported to have found 2 items in a 14-item test severely biased by use of a test of fit to the one-parameter ICC model. However, two minority group reviewers had objected to the items without any knowledge of the statistical analysis. In the latter, Subkoviak and his colleagues inserted into a widely-used college admissions test, the *College Qualification Test* (CQT) items from the *Black Intelligence Test of Cultural Homogeneity* (BITCH). These items are designed to be biased in favor of members of a black culture. The three-parameter ICC technique, using areas between the curves as the measure of discrepancy, was best at detecting that the BITCH items were biased, but it is unlikely that any test developer sensitive to the issues of item bias would have mixed items from BITCH with items from CQT in a serious attempt to build a useful test for college admission. It requires no statistical microscope aiding the naked eye to observe that items as extreme as these do not belong in the instruments in which they were mixed.

However, there are also cases in which items have been branded as biased by judges, but statistical analysis reveals no presence of bias. Perhaps the most flagrant is the case of the sixth item in the WISC-R comprehension test about boys fighting. It has been vehemently contended that the keyed answer is biased against blacks. However, Sandoval (1979) found no difference between blacks and whites on this item, and Cotter and Berk (Note 57) found the item slightly biased in favor of blacks. Jensen (1976) and Sandoval and Miille (1980) found that their judges were no more successful than would be expected by chance in detecting which items in the *Wonderlic Personnel Test* and *WISC-R* would be more difficult for blacks than whites. Jensen used psychologists as judges, Sandoval and Miille used college students, and both investigators included judges of both races. Cotter and Berk (1957) reported that only a relatively low percentage of items flagged by judges as biased were found to be discrepant by statistical analysis. Hills and King (Note 42) similarly found that judges given both discrepant and concordant items to evaluate for bias were unable to sort them with reasonable accuracy and were unable to give reasons for bias that were in agreement.

Part of what happens when judges attempt to pick the items that are biased in a test seems to be that they overreact to problems they think the examinee will have but does not have. Black children apparently know how to give the keyed answer to the question about the boys fighting, though that may not be what they would do in a real life situation. Similarly, it was expected by blacks that the question on Florida's SSAT-II which asks about filling out a check would be outside the range of experience of black children, but statistically blacks performed as well or better on that item than did whites. The judges don't give the children or their teachers the credit they deserve.

On the other hand, the statistical methods may flag as discrepant items which reveal existing differences between groups in the content being tested. Such items are not biased, of course, since only differences in performance due to irrelevant factors are included in our definition of bias. If items reveal actual differences in what we are trying to measure, they are doing just what good items should do, and they should not be removed from our tests. Thus, Rudner (Note 3) found the hearing-impaired seemed to find negatives and comparatives particularly difficult. But since these are relevant, i. e., they are part of what the test should be measuring, Rudner concluded that teachers and curriculum developers should pay particular attention to these concepts when working with the hearing-impaired. Similarly, Linn and Harnisch (1981) found that blacks had more trouble on items concerning the metric system, but less trouble than whites on story items dealing with money. However, since both contents are appropriate for the test in which they were included, this can hardly be called bias, and Linn and Harnisch did not recommend that the items be excluded from the test.

Jensen (1976) and Miele (1979) point out another factor which may cause statistical procedures to flag items but which can not reasonably be included as bias. They found that some items seem merely to detect those people whose development is not as far along as others. The apparent "bias" can be removed by comparing children from one group with children who are two years older from the other group. Items which perform this way are telling us something important and useful about the two groups; such items are not misleading us through distortion based on an irrelevant element. Rock and Werts (Note 19) similarly attribute some instances of differential difficulty of items to differences in development, not to bias in the items. Donlon, Hicks, and Wallmark (1980) found that the scores of girls on math on the Graduate Record Examination Aptitude Test could be improved by replacing items on graphs with items on algebra, but they did not recommend doing so since both algebra and graphs are legitimate content for the mathematics part of the test.

Ironson and Subkoviak (1979) noted another factor which is, indeed, irrelevant, but is not a property of the item. That is the location of the item in the test. They found that the statistical procedures flagged items near the end as biased, but that may only reflect the test-taking behavior of different cultural groups, and no change in an item or removal of an item will keep some items from being near the end of the test. Similarly, Bode (Note 54) noted that when items favored the Hispanics in her study, the white sample had many more omits than did the Hispanic sample. This may be a difference in a test-taking characteristic of the cultures which is irrelevant to the content of the item and cannot be corrected by replacing a particular item with another of similar characteristics without the "biasing" factor. According to Bode's analysis, one could not solve this problem without changing the difficulty level of the test, a far cry from the problem of item bias.

The fact that statistical analyses often flag as discrepant and potentially-biased items that are perfectly sound and should not be removed or replaced with other items from the same pool, does not mean that this is always the case. A few patterns that reflect irrelevancies have been found. In one of the earliest studies, Coffman (1961) found that men were favored by SAT-V items with mechanical or business content, and women were favored by items with personal or personality content. Donlon (Note 56) found that same effect in SAT-V ten years later. This may be a genuine irrelevant factor which could be considered item bias in a verbal aptitude test.

Statistical analyses of one kind may flag as discrepant items that a different kind of analysis would not flag. The various approaches to statistical detection of discrepant items are not based on the same concept of what about an item's statistical behavior would indicate bias, they do not make the same assumptions, and they differ in statistical elegance. Thus, while it is true that a statistical flag does not mean that an item is biased, it is equally true that items which escape being flagged by any particular statistical technique, or set of techniques, are not necessarily free from bias. Using another statistical technique, or even a modest modification of a particular technique (see Raju, Note 53, and Bode, Note 54, for example) may result in different items being flagged as discrepant. Unless one used every conceived and conceivable technique and an item escaped being discrepant on all of them, one could not be sure that a particular item should not be examined for bias by judges. Using the technology of the present day for flagging potentially-biased items with statistics, an item cannot get a clean bill of health.

Clearly, then, though flagrantly-biased items are easily detected "with the naked eye," items whose bias is not so flagrant cannot easily be detected statistically or by judges, and, in fact, cannot surely be detected or absolved of guilt by use of both statistical methods and expert judges. However, when a statistical technique points out an item which may be biased, and when judges can detect the irrelevant factor which produces the bias, the artifact and its results may reappear regularly over a period of years.

The third conclusion that we reached earlier was that ordinarily when biased items are located and the source of their bias is clear, removal of the biased items and replacement of them by other items from the same pool but without the biasing factor does not remove the differences between group means that caused us to be concerned about item bias in the first place (Coffman, 1978; Flaugher & Schrader, Note 23; Green, 1972; Linn et al., 1980; Ozenne, in Bianchini, 1977). In some cases the mean differences may not change because the amount of bias favoring one group is counterbalanced by a similar amount of bias favoring the other group (Coffman, 1978). In one case, the amount of difference in mean scores could be appreciably reduced by concentrating all the SAT-M items on algebra. Since women are less handicapped in algebra than in other kinds of mathematics items, the difference between their performance and the average performance of men on the SAT-M would be reduced if other kinds of mathematics items were avoided. However, the content validity of the test would be altered by such a modification. The test would no longer be a measure of mathematics aptitude;it would become a measure of algebra achievement (Donlon, Note 56).

In the cases of both Coffman's study and Donlon's study, one in which the mean difference could be removed by counterbalancing items, and one in which the mean difference could be reduced by changing the nature of the test, the groups being compared were males and females. It appears that when mean differences are small, they can sometimes be removed by balancing out items favoring one group with items favoring the other group, or even by removing items which appear to be biased. But small differences in group means usually do not involve ethnic groups who are concerned about item bias. Small differences are more likely to involve comparisons between males and females or between culturally similar groups who have studied the same subject using different curricular approaches. The possibility of removing mean differences by selecting items carefully from within an item pool can be evaluated using the GD/I ratio introduced by

Jensen (1980a). At least one case has been noted in the literature in which removing the items that appeared to be biased would increase the difference between group means instead of decreasing it (Arneklev, Note 52).

If the above conclusions are found to be supported by studies in the future, what value is there in studying item bias at all? If (a) judges can and do readily remove flagrantly biased items so that they never appear in tests at all; (b) statistical tests to detect potentially-biased items do not agree with each other so that no one technique or combination of techniques can be trusted to clearly pinpoint biased items or absolve items of the possibility of bias; (c) judges cannot be trusted to detect subtle bias in items that would be detectable if statistical analyses were used; and (d) even using both statistical techniques and expert judges to locate biased items does not remove the mean differences that aroused the concern--and may even increase them, is there any point to spending the time and effort required to check on the possibility of biased items appearing in tests? There may be.

Judging from the history of the study of item bias that has been reported here, the most beneficial use of item-bias search techniques may be to rescue for the item pool items that judges think are biased but which turn out not to be biased at all. It seems highly likely that our judges or item writers who discard as potentially-biased item ideas before ever developing them to tryout level may be throwing out babies thinking they are bath water. Those items may be like the WISC-R item on boys fighting or the SSAT-II item on check writing, perfectly useful though appearing to the hypersensitive reviewer, editor, or critic to be obviously biased.

This would imply that item bias studies should take place early in the test development activity, long before the final test is assembled. It would also imply that the item writers should go ahead and produce relevant items even though there might be a strong suspicion that the items might turn out to be biased; those strong suspicions are often incorrect.

Another possible use for item-bias research would be for tests in contents that have not previously encountered examinees from particular cultural groups. For example, if women have not traditionally applied for positions as fire fighters, and those who develop tests for fire fighters have never had to think about whether test items might be biased against women, it may be that the first women who apply for the positions or the training for it will find the items entirely male-oriented. The same might apply to other positions which have traditionally been held by people from one cultural group—or not held by people from a particular cultural group. The armed forces, nursing, coaches of female athletic teams, airline pilots, police, race-car drivers, jockeys, sea captains, harbor pilots, secretaries, receptionists, plumbers, building contractors, crane operators, ships' radio officers, are just a few examples of occupations in which item bias may never have been seriously considered simply because it never came up as an issue until lately. Item-bias statistical methods, followed by evaluations of potentially-biased items by judges who determine whether the apparent biasing factor can be isolated as irrelevant, may indeed be helpful for examinations for such occupations, to demonstrate that despite lack of attention bias is not a problem, to demonstrate that item pools must be modified to remove items contaminated by irrelevant factors, and to rescue items that only appear to be biased.

Reference Notes

1. Angoff, W. H. *The investigation of test bias in the absence of an outside criterion.* Paper presented at National Institute for Education Conference on Test Bias, December 4, 1975. (mimeo)
2. Breland, H. M., Stocking, M., Pinchak, B. M., & Abrams, N. The cross-cultural stability of mental test items: An investigation of response patterns for ten socio-cultural groups. *Project Report* 74-2. Princeton, N. J.: Educational Testing Service, 1974.
3. Rudner, L. M. *An approach to biased item identification using latent trait measurement theory*. Paper presented to the American Educational Research Association, 1977. a.
4. Scheuneman, J. *A method of assessing bias in test items in the absence of an outside criterion*. New York: Psychological Corporation, undated, mimeo, circa 1977.
5. Sinnott, L. T. *Differences in item performance across groups*. Research Report 80-19. Princeton, N. J.: Educational Testing Service, 1980.
6. Scheuneman, J. *A new method of assessing bias in test items.* Paper presented to American Educational Research Association, Washington, 1975.
7. Groome, M. L. & Groome, W. R. *Item bias identification: A comparison of two approaches*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, 1979.
8. Scheuneman, J. *Validating a procedure for assessing bias in test items in the absence of an outside criterion*. Paper presented to American Educational Research Association, 1976.

9. Angoff, W. H. *General considerations in the study of test bias when the criterion is absent.* Paper presented to American Educational Research Association, 1976.

10. Pine, S. M. & Weiss, D. J. Effects of item characteristics on test fairness. *Research Report 76-5*. Minneapolis: University of Minnesota, 1976.

11. Nungester, R. J. *An empirical examination of three models of item bias*. Unpublished dissertation. Florida State University, 1977.

12. Angoff, W. H. & Stern, J. The equating of the scales for the Canadian and American *Scholastic Aptitude Tests*. *CEEB Research and Development Reports* 71-72, No. 4. New York: College Entrance Examination Board, 1971.

13. Strassberg-Rosenberg, B. & Donlon, T. F. *Content influences on sex differences in performance on aptitude tests*. Paper presented to National Council on Measurement in Education, 1975.

14. Hicks, M. M., Donlon, T. F., & Wallmark, M. M. *Sex differences in item responses on the Graduate Record Examination*. Paper presented to the National Council on Measurement in Education, 1976.

15. Hunter, J. E. *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items*. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD, 1975.

16. Angoff, W. H. *The use of difficulty and discrimination indices in the identification of biased test items*. Paper presented to National Symposium on Educational Research, Johns Hopkins University, 1980.

17. Rudner, L. M. & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented to the American Educational Research Association, 1978.

18. Wright, B. D., Mead, R. & Draba, R. Detecting and correcting test item bias with a logistic response model. Research Memorandum No. 22. Statistical Laboratory, College of Education, University of Chicago, October 1976.

19. Rock, D. A. & Werts, C. E. Construct validity of the SAT across populations--an empirical confirmatory study. CEEB Research and Development Report 79-2. Princeton, N. J.: Educational Testing Service, 1979.

20. Cardall, C. & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. Research Bulletin, 64-61. Princeton, N. J.: Educational Testing Service, 1964.

21. Fishbein, R. L. An investigation of the fairness of the items of a test battery. Unpublished dissertation, Florida State University, 1974.

22. Rudner, L. M. Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977. b.

23. Flaugher, R. L. & Schrader, W. B. Eliminating differentially difficult items as an approach to test bias. Research Bulletin 78-4. Princeton, N. J.: Educational Testing Service, 1978.

24. Ironson, G. H. Chi-square and latent trait approaches to the measurement of item bias. Paper presented at National Symposium on Educational Research, Johns Hopkins University, 1980.

25. Veale, J. R. & Foreman, D. I. Cultural validity of items and tests: A new approach, *CORE Technical Report*, 1975, unpublished, mimeo.

26. Veale, J. R. & Foreman, D. I. Cultural variations in items and tests: A global analysis. *REAL Technical Report No. 1*, 1976, unpublished, mimeo.

27. Kohr, R. L. & Tung, P. S. *Cultural validity, Main program*. University Park, PA: Penn State Computer Center, 1977.

28. Veale, J. R. Estimation of cultural bias and true ability levels using p-values test scores, and foil response data. *Statistical Research Technical Report No. 1*, 1976, unpublished, mimeo.

29. Petersen, N. S. *Bias in the selection rule--bias in the test*. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977.

30. Veale, J. R. A theoretical framework for the cultural variation procedures and a new paradigm for cultural validation. *Statistical Research Technical Report No. 5*, 1978, unpublished, mimeo.

31. Shepard, L. A. *Definitions of bias.* Paper presented to National Symposium on Educational Research, Washington, 1980.

32. Divgi, D. R. *Does the Rasch model really work? Not if you look closely.* Paper presented at Annual Meeting of the National Council on Measurement in Education, 1981(a).

33. Gustafson, J. -E. *PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items.* Report 85. Institute of Education, University of Goteborg, 1979.

34. Durovic, J. J. Test bias: An objective definition for test items. Paper presented to Northeastern Educational Research Association, 1975.

35. Draba, R. E. The identification and interpretation of item bias. Research Memorandum No. 26. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1977.

36. Divgi, D.R. Potential pitfalls in applications of item response theory. Discussant's remarks at Annual Meeting of the National Council on Measurement in Education. 1981(b).

37. Carlton, S. T. & Marco, G. L. Methods used by Educational Testing Service testing programs for detecting and eliminating item bias. Paper presented at National Symposium on Educational Research, Johns Hopkins University, 1980.

38. Green, D. R. Reducing bias in achievement tests. Paper presented at the annual meeting of the American Educational Research Association, 1976.

39. Merz, W. R. & Grossen, N. E. An empirical investigation of six methods of examining test item bias. Paper presented at the annual meeting of the National Council on Measurement in Education, 1979.

40. Merz, W. R. & Grossen, N. E. An empirical investigation of six methods for examining test item bias. Final Report to National Institute for Education, Grant NIE-6-78-0067, undated.

41. Ozenne, D. G., Van Gelder, N. C., and Cohen, A. J. Achievement test restandardization: Emergency School Aid Act National Evaluation. Santa Monica, CA: System Development Corporation, 1974.

42. Hills, J. R. & King, F J *Evaluation of a method for detecting potentially biased items in the FLORIDA STATEWIDE ASSESSMENT TESTS.* Tallahassee, FL: Student Assessment Section, Department of Education, 1981.

43. Sowell, T. *Assumptions versus history in ethnic education.* Paper presented to American Educational Research Association, 1981.

44. Urry, V. W. *A Monte Carlo investigation of logistic mental test models.* Unpublished doctoral dissertation, Purdue University, 1970.

45. Wingersky, M. *Test whether the item parameters estimated by LOGIST for two separate groups differ significantly.* Draft manuscript, Educational Testing Service, 1977.

46. Scheuneman, J. *Consistency across administration of certain indices of bias in test items.* Paper presented to American Educational Research Association, 1980.

47. Subkoviak, M. J., Mack, J. S., & Ironson, G. H. *Item bias detection procedures: empirical validation.* Paper presented at annual meeting of American Educational Research Association, 1981.

48. Douglass, J. B. *Item bias, test speededness, and Rasch tests of fit.* Paper presented at annual meeting of American Educational Research Association, 1981.

49. Wright, B. D., Mead, R., & Bell, S. BICAL: calibrating items with the Rasch model. *Research Memorandum No. 23B.* Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1979.

50. Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters.* (Research Memorandum 76-6). Princeton, N. J.: Educational Testing Service, 1976.

51. Frary, R. B. & Giles, M. B. *Multiple-choice test bias as reflected by examinee selection of inappropriate answers*. Report submitted to National Institute of Education, 1980.

52. Arneklev, B. L. *Data related to the question of bias in standardized testing*. Tacoma, WA: Office of Evaluation, Tacoma Public Schools, 1975.

53. Raju, M. S. *A review of item bias detection procedures employed with the SRA Achievement Series*. Paper presented at National Symposium on Educational Research, Johns Hopkins University 1980.

54. Bode, R. K. *Comparison of pretest and reanalysis results of an item bias study*. Paper presented at the annual meeting of the American Educational Research Association, 1981.

55. Angoff, W. H. & Sharon, A. T. Patterns of test and item difficulty for six foreign language groups on the *Test of English as a Foreign Language*. *CEEB Research and Development Reports* 71-72, No. 5. New York: College Entrance Examination Board, 1972.

56. Donlon, T. *Content factors in sex differences on test questions*. Paper presented at 1971 meeting of New England Educational Research Organization, 1971.

57. Cotter, D. E. & Berk, R. A. *Item bias in the WISC-R using black, white, and Hispanic learning disabled children*. Paper presented at the annual meeting of the American Educational Research Association, 1981.

References

Angoff, W. H. & Ford, S. H. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, 10, 95-105.

Angoff, W. H., & Herring, C. L. Study of the appropriateness of the Law School Admissions Test for Canadian and American Students. Report No. LSAC-71-1. In Law School Admissions Council, *Reports of LSAC Sponsored Research: Volume II, 1970-1974*. Princeton, N. J.: Law School Admissions Council, 1976.

Baker, F. B. A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement*, 1981, 18, 59-61.

Bianchini, J. C. Achievement tests and differential norms. In M. J. Wargo & D. R. Green (Eds.) *Achievement testing of disadvantaged and minority students for educational program evaluation*. Monterey, CA: CTB/McGraw-Hill, 1977, pp. 157-180.

Bianchini, J. C., & Loret, P. G. *Anchor test study final report. Project report and volumes 1 through 30 and anchor test study supplement. Volumes 31 through 33*. Berkeley CA: Educational Testing Service, 1974. (ERIC Document Reproduction Service Nos. ED 092 601 through ED 092 634).

Cleary, T. A. & Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, 28, 61-75.

Coffman, W. E. Sex differences in responses to items in an aptitude test. *Eighteenth Yearbook of National Council of Measurement in Education*, 1961, 117-124.

Coffman, W. E. *An exploratory study of differences in the performance of black and white pupils in grades 6, 7, 8 and 9 on the items in the Iowa Tests of Basic Skills*. Washington, D. C.: Educational Resources Information Center (ERIC), ED 156713, 1978.

Donlon, T. F., Hicks, M. M., & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. *Applied Psychological Measurement,* 1980, 4, 9-20.

Echternacht, G. A quick method for determining test bias. *Educational and Psychological Measurement*, 1974, 34, 271-280.

Good, I. J. The surprise index for the multivariate normal distribution. *The Annals of Mathematical Statistics*, 1956, 27(4), 1130-1135.

Green, D. R. *Racial and ethnic bias in test construction*. Monterey, CA: CTB/McGraw-Hill, 1972.

Green, D. R. What does it mean to say a test is biased? *Education and Urban Society*, 1975, 8, 33-52.

Green, D. R. & Draper, J. R. *Exploratory studies of bias in achievement tests*. Monterey, CA: CTB/McGraw-Hill, 1972.

Hilton, T. L. & Rhett, H. *The base-year survey of the National Longitudinal Study of the high school class of 1972*. Princeton, N. J.: Educational Testing Service, 1973.

Hoepfner, R. Achievement test selection for program evaluation. In J. J. Wargo & D. R. Green (Eds.) *Achievement testing of disadvantaged and minority students for educational program evaluation*. Monterey, CA: CTB/McGraw-Hill, 1977, pp. 277-323.

Ironson, G. H. & Subkoviak, M. J. A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 1979, 16, 209-225.

Jensen, A. R. Test bias and construct validity. *Phi Delta Kappan*, 1976, 58, 340-346.

Jensen, A. R. *Bias in mental testing*. New York: Free Press, 1980, a.

Jensen, A. R. Uses of sibling data in educational and psychological research. *American Educational Research Journal*, 1980, 17, 153-170, b.

Linn, R. L. & Harnisch, D. L. Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 1981, 18, 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. *An investigation of item bias in a test of reading comprehension.* Washington, D. C.: Educational Resources Information Center (ERIC), ED-184091, 1980.

Lord, F. M. A study of item bias, using item characteristic curve theory. In International Congress of the International Association for Cross-Cultural Psychology. *Basic problems in cross-cultural psychology*. Amsterdam: Swetz & Zeitlinger, B. U., 1977, 19-29, a.

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, 14, 117-138, b.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, N. J.: Erlbaum, 1980.

Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores*. Reading: Addison-Wesley, 1968.

Marascuilo, L. A. & Slaughter, R. E. Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement*, 1981, 18, 229-248.

Merz, W. R. Test fairness and test bias: A review of procedures. In M. J. Wargo & D. R. Green (Eds.) *Achievement testing of disadvantaged and minority students for educational program evaluation*. Monterey, CA: CTB/McGraw-Hill, 1977, pp. 129-152.

Miele, F. Cultural bias in the WISC. *Intelligence*, 1979, 3, 149-164.

Petersen, N. S. & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, 13, 3-29.

Plake, B. S. & Hoover, H. D. An analytical method of identifying biased test items. *Journal of Experimental Education*, 1979/80, 48 (2), 153-154. Pyrczak, F. Passage-dependence of items designed to measure the ability to identify the main ideas of paragraphs: Implications for validity. *Educational and Psychologial Measurement*, 1974, 34, 343-348.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 1979, 4, 207-230.

Rudner, L. M. Using standard tests with the hearing impaired: The problem of item bias. *The Volta Review*, 1978, 80, 31-40.

Rudner, L. M., Getson, P. R., & Knight, D. L. A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, 1980, 17, 1-10, a.

Rudner, L. M., Getson, P. R., & Knight, D. L. Biased item detection techniques. *Journal of Educational Statistics*, 1980, 5, 213-233, b.

Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, 1979, 47, 919-927.

Sandoval, J. & Miille, M. P. W. Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 1980, 48, 249-253.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Scheuneman, J. Latent trait theory and item bias. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.) Psychometrics for educational debates. New York: Wiley, 1980.

Scheuneman, J. D. A response to Baker's criticism. Journal of Educational Measurement, 1981, 18, 63-66.

Shepard, L., Camilli, G., & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Journal of Educational Statistics, 1981, 6, 317-375.

Swineford, F. Comparisons of black candidates and Chicano candidates with white candidates. Report No. LSAC-72-6. In Law School Admissions Council, Reports of LSAC Sponsored Research: Volume II, 1970-1974. Princeton, N. J.: Law School Admissions Council, 1976.

Test bias fears are exaggerated, Ebel says. Report on Educational Research, December 17. Washington, D. C.: Capitol Publications, 1975, pp. 7-8.

Tinsley, H. E. A. & Dawis, R. V. An investigation of the Rasch simple logistic model: Sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Tucker, L. R An interbattery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Tuinman, J. J. Determining the passage dependency of comprehension questions in 5 major tests. Reading Research Quarterly, 1973-4, 9, 206-223.

Wargo, M. J. & Green, D. R. (Eds.) Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill, 1977.

Wightman, L. E. Law School Admission Test: Comparisons of Canadian candidates with non-Canadian candidates. Report No. LSAC-76-5. In Law School Admissions Council, Reports of LSAC Sponsored Research: Volume III, 1975-1977. Princeton, N. J.: Law School Admissions Council, 1976.

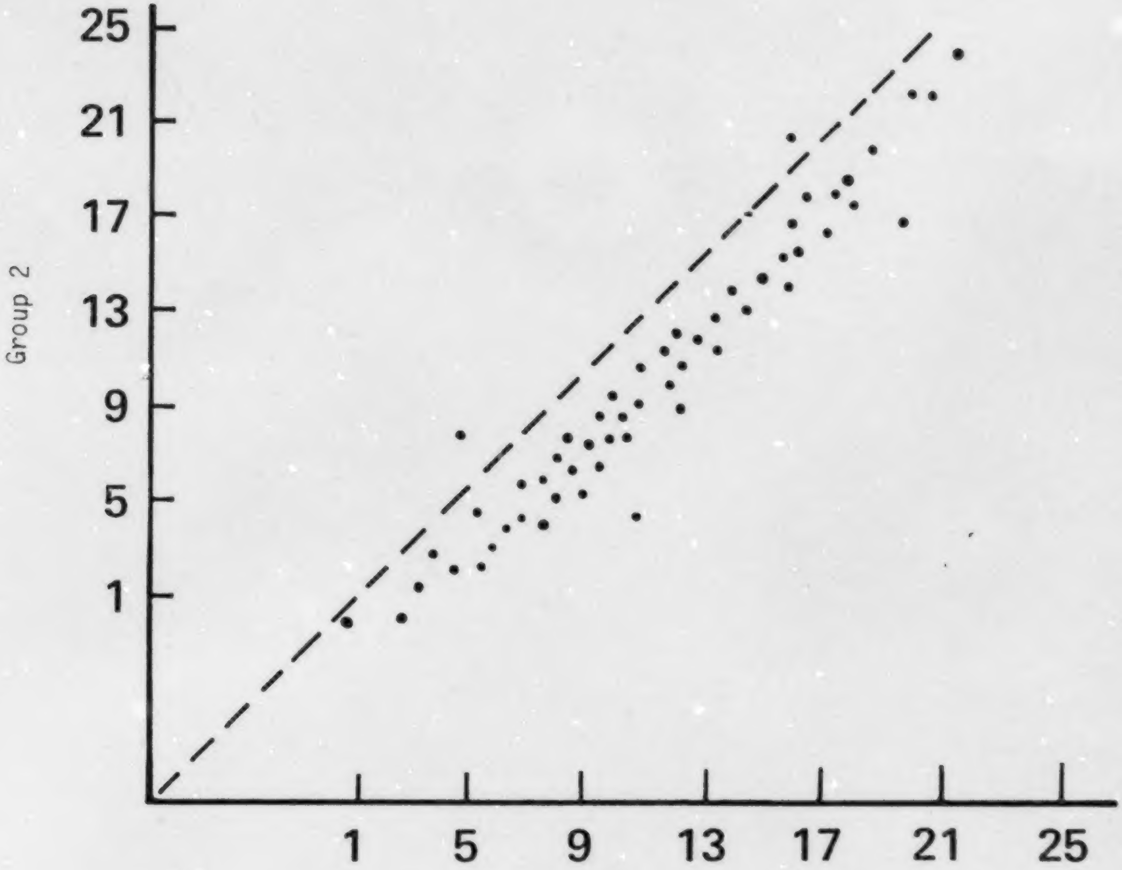
Williams, R. I. Black pride, academic relevance & individual achievement. The Counseling Psychologist, 1970, 2, 18-22.

Williams, R. I. Abuses and misuses in testing black children. The Counseling Psychologist, 1971, 2, 62-73.

Willingham, W. W. GRE validity and the Graduate Record Examination Program. Princeton: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.



Group 1

Figure 1

Item easiness values for two groups

Delta scale, Mean=13, S.D.=4

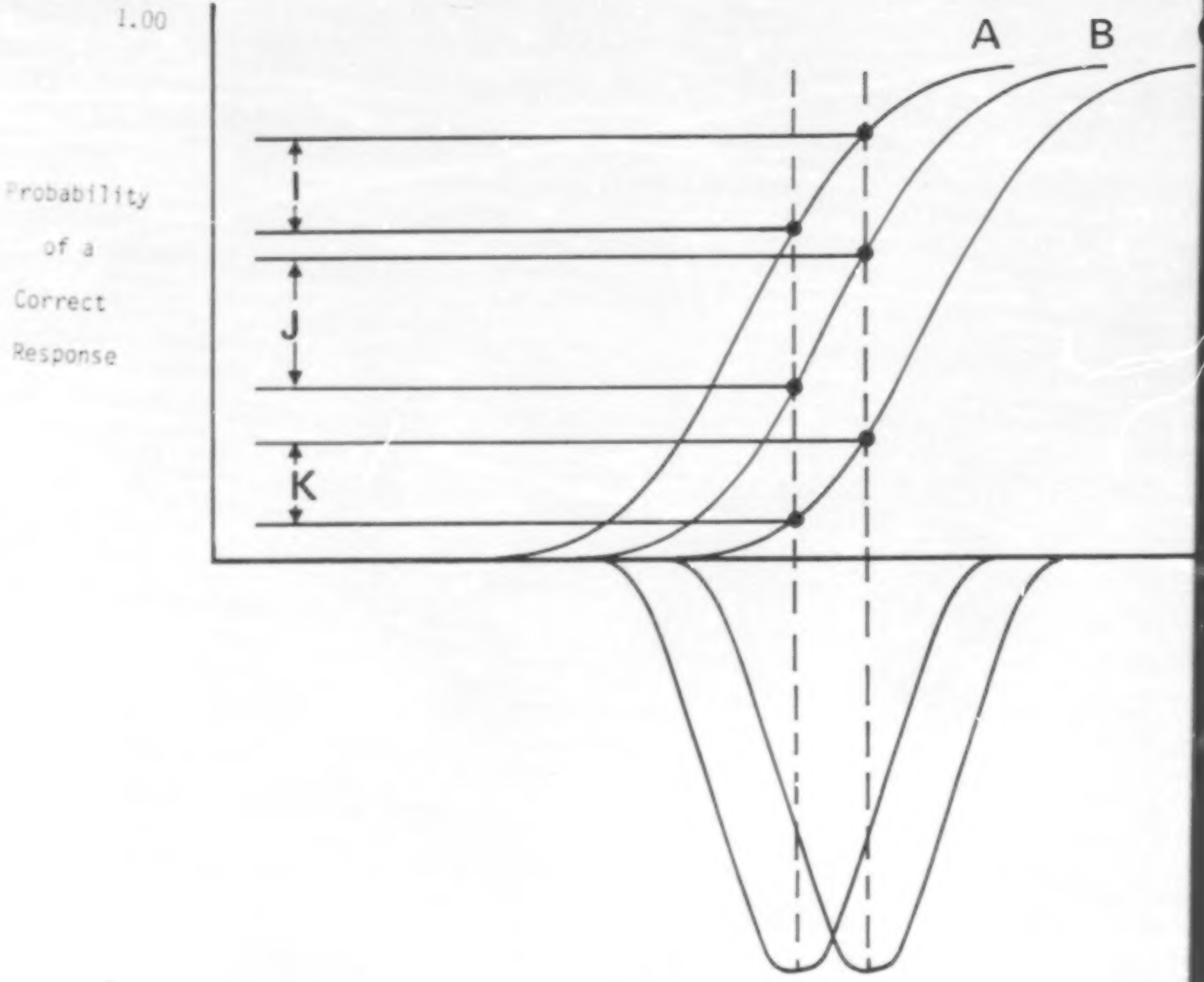


Figure 2

Item-group interaction from unbiased items of equal discrimination when groups differ in mean ability. (Based on Hunter, 1975)